(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
08 October 2020 (08.10.2020)

WIPO | PCT

(10) International Publication Number
WO 2020/205649 A1

(51) International Patent Classification:
G16B 40/20 (2019.01) *G16H 50/20* (2018.01)
G16H 10/40 (2018.01) *G16H 50/70* (2018.01)

(21) International Application Number:
PCT/US2020/025502

(22) International Filing Date:
27 March 2020 (27.03.2020)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/826,228 29 March 2019 (29.03.2019) US

(71) Applicant: **VENN BIOSCIENCES CORPORATION** [US/US]; Two Tower place, 5th Floor, South San Francisco, California 94080 (US).

(72) Inventors: **SERIE, Daniel**; c/o Venn Biosciences Corporation, Two Tower place, 5th Floor, South San Francisco, California 94080 (US). **WU, Zhenqin**; c/o Venn Biosciences Corporation, Two Tower place, 5th Floor, South San Francisco, California 94080 (US).

(74) Agent: **OSTOMEL, Todd** et al.; Squire Patton Boggs (US) LLP, 275 Battery Street, Suite 2600, San Francisco, California 94111 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) Title: AUTOMATED DETECTION OF BOUNDARIES IN MASS SPECTROMETRY DATA

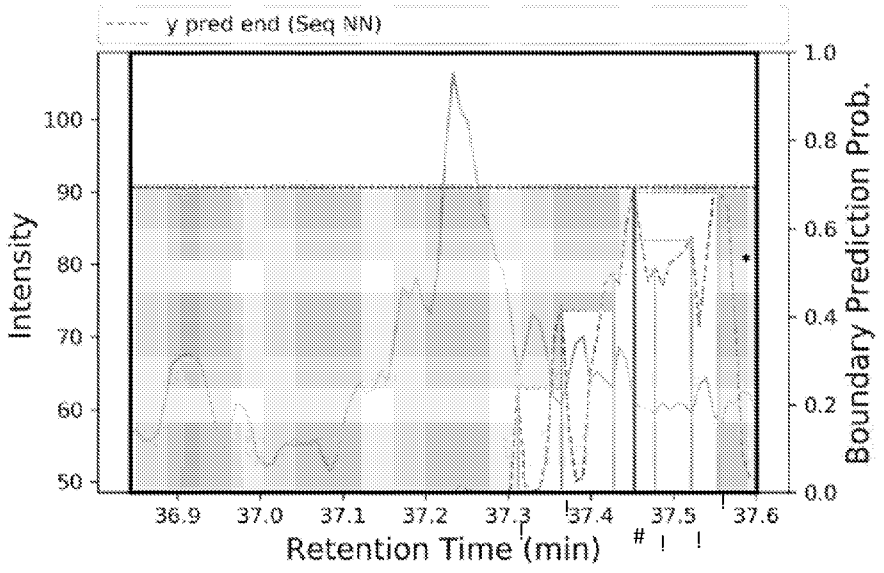


FIG. 12

(57) Abstract: A system and method for automated detection of the presence or absence of a quantity based on intensities expressed in terms of, or derived from frequency or time dependent data. According to one example intensities from mass spectrometry are identified using a non-linear mathematical model, such as an artificial neural network trained to find start and stop peaks of an intensity, from which an abundance may be determined.

WO 2020/205649 A1

[Continued on next page]

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:

— *with international search report (Art. 21(3))*

AUTOMATED DETECTION OF BOUNDARIES IN MASS SPECTROMETRY DATA

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to, and the benefit, of US Provisional Patent Application No. 62/826,228, filed March 29, 2019, the entire contents of which are herein incorporated by reference in its entirety for all purposes.

FIELD

[0002] This disclosure is generally related to detection of physical things using mass spectrometry and, in particular, the automated detection of a biological quantity using multiple reaction monitoring mass spectrometry (MRM-MS).

BACKGROUND

[0003] Multiple reaction monitoring (MRM) is a technique utilized in tandem mass spectrometry (MS) which allows for highly sensitive and specific detection of proteins, lipids, and post-translational modifications (PTMs), among other analytes. This technology holds great promise for use in new clinical assays and diagnostics, but the availability of precise analytical software remains a bottleneck.

[0004] Traditional analysis of MRM experiments involved choosing the start and stop of a chromatographic peak by hand, working from transitions previously characterized by data-dependent acquisition or other discovery techniques. A typical implementation can be seen in Agilent's MassHunter software, where intensities are plotted over a pre-specified range of retention times (RT) for a given precursor and product mass-to-charge ratio. Selection of the beginning and end of the peak yields an observed RT, peak width, and integrated abundance value. Person to person variability in assessment, human error, and a large time investment render this method inadequate for non-research use. In recent years many software packages have attempted to fill this gap by automatically integrating peaks from a list of transitions.

[0005] The existing software packages have proven very useful in quantifying, e.g., peptides, when there is a high relative abundance present in the mass spectrometry data. But these software packages are less accurate in integrating relatively low-abundant and highly-heterogeneous species such as glycosylated peptides.

SUMMARY

[0006] The instant disclosure sets forth methods and systems for automated detection of the presence of a quantity expressed in terms of, or derived from time series data. The instant

disclosure also sets forth methods and systems for automated detection of the presence of a quantity expressed in terms of, or derived from frequency-dependent data.

[0007] In one implementation, automated detection is applied to a time series of intensities vs. time. The intensities derive from a detector which detects the presence or absence of a quantity. For example, the automated detection may be applied to mass spectrometry (MS) in which the intensities derive from a MS detector detecting a transition value as a function of time, *e.g.*, retention time. As such, the automated detection described herein may be applied to a wide variety of data inputs based on intensity values observed as a function of time or as a function of frequency.

[0008] In one implementation the automated detection is applied to a time series of mass to charge ratio intensities vs. time, which were produced from a process involving mass spectrometry (MS). In another implementation the automated detection is applied to a time series of intensities vs. time, which were produced from a process involving liquid or gas chromatography.

[0009] According to other examples the automated detection may be applied to any of the following intensities expressed in terms of, or derived from time or frequency data:

- x-ray diffraction in which the intensities derive from an x-ray detector, *e.g.*, charge coupled device (CCD), detecting a reflection of x-rays as a function of the x-ray scan angle, *e.g.*, 2-Θ, in which scan angles are observed as a function of time.
- Raman spectroscopy in which the intensities derive from an infrared detector detecting vibrational modes as a function of light wavelength, *e.g.*, cm^{-1}, absorption or emission.
- UV-VIS spectroscopy in which the intensities derive from a light detector detecting intensities of electronic transitions as a function of ultra-violet and visible light wavelengths or frequencies. fluorimetry spectroscopy or phosphorimetry spectroscopy in which emitted fluorescence and phosphorescence light intensity is detected as a function of incident, *i.e.*, absorbed or excitation, wavelength.
- Nuclear Magnetic Resonance (NMR) spectroscopy in which the intensities derive from a radio wave detector detecting nuclear transitions as a function of radiation frequency absorption or emission within a strong magnetic field. In NMR, the intensity v. frequency detection can be converted, *e.g.*, using Fourier transformations, into an equivalent intensity v. time. The same principles applicable to NMR are also applicable to Magnetic Resonance Imaging (MRI).

[00010] In one particular implementation automated detection is applied to a time series of mass to charge ratio intensities vs. time, which were produced from a process involving mass spectrometry. According to this implementation the automated detection involves using one or more mathematical models to detect where a start peak and end peak occurs in MS data. By providing an accurate selection of these boundary points a more accurate assessment of an abundance is made.

[00011] According to one aspect, a method of training a machine learning model to identify mass spectrometry peaks uses as a training set sequences of points representing mass/charge ratios where the peak starts and peak stops are labeled. The trained model is adapted for identifying mass spectrometry start and stop peaks indicating the presence of one or more glycopeptides or peptides, or a fragment thereof, present in a biologic sample.

[00012] According to another aspect, a method of identifying mass spectrometry peaks uses at least one trained neural network model that produces an output as a function of input parameters, wherein values for the input parameters may be a sequence of retention time values for a glycopeptide having a m/z (mass to charge) ratio.

[00013] According to another aspect, raw MS data is featurized to produce training sets for a machine learning model for purposes of training the model to identify peak start and stop times in MS data. The featurization may include discretizing an analog sample of MS data into a sequence of points, wherein retention times corresponding to peak areas in the analog sample are left out of the sequence of points, and assigning labels to points among the sequence of points corresponding to peak start and peak stop times to produce a labeled sequence of points. The labels may be found using human annotators trained to select peak start and stop times using a display and input device.

[00014] According to another aspect, a method computes a distribution of probabilities over a sequence of points representing MS data for a given m/z ratio and retention time period, the probabilities representing a likely peak start or peak stop retention time. From these distribution of probabilities a statistic for an abundance is computed. The statistic may be a mean.

[00015] According to another aspect, a method of diagnosing a disease includes quantifying mass spectrometry transitions, wherein the transitions are m/z intensity value peaks characteristic of glycopeptides or peptides, or a fragment thereof, comprising an amino acid sequence and analyzing the quantification using a model to determine a probability that the sample is or isn't within a disease classification based on the quantification. In some examples, the glycopeptides or peptides, or a fragment thereof, comprising an amino acid

sequence are selected from those sequences set forth in International Patent Application Publication No. WO 2019/046814 - Identification and use of glycopeptides as biomarkers for diagnosis and treatment monitoring; International Patent Application Publication No. WO2019079639A1 - Identification and use of biological parameters for diagnosis and treatment monitoring; US Patent Application Publication No. US20190101544A1 - Identification and use of glycopeptides as biomarkers for diagnosis and treatment monitoring; International PCT Patent Application No. PCT/US2020/0162861, filed January 31, 2020; and also U.S. Provisional Patent Application No. 62/989,510, filed March 13, 2020. The entire contents of each of these patent applications is herein incorporated by reference in its entirety for all purposes.

BRIEF DESCRIPTION OF THE DRAWINGS

[00016] FIG. 1 shows a process for calculating an abundance from mass-spectrometry data using a machine learning model.

[00017] FIG. 2 shows a process for training the machine learning model of FIG. 1. Model performances on test set. **A**: Bar plot of absolute error in boundary predictions on test set. **B**: Cumulative distributions of boundary prediction errors. Note that 19039 (28.1%) samples are out of range (> 6 seconds) in rule-based predictions, 4824 (7.1%) samples in Sequential NN, 1470 (2.2%) samples in Reference-based NN. C: Scatter plot of peak abundance prediction/annotation on the test set.

[00018] FIG. 3 shows a process for generating a sequence of points from MS sample used as input to the machine learning model of FIG. 2.

[00019] FIG. 4 is a process for computing an abundance based on computed probabilities of peak start and peak stop retention times in an MS sample.

[00020] FIGS. 5A and 5B show a model architecture for a first machine learning model (MODEL A) that is a bi-directional recurrent neural network with long short term memory units and attention layer.

[00021] FIGS. 6A and 6B show a model architecture for a second machine learning model (MODEL B) that is a bi-directional recurrent neural network with long short term memory units and attention layer. FIG. 6C shows an attention map for a sample using MODEL B. In FIG. 6C, the sample peak and reference-based model predictions are for the MRM transition for SEQ ID NO:1 in International PCT Patent Application No. PCT/US2020/0162861, filed January 31, 2020, which is herein incorporated by reference in its entirety for all purposes, and its reference peak illustrated in lower right panel, and

attention map (extracted from the model) in lower left panel.

[00022] FIGS. 7A and 7B shows sample labeled input and computed probabilities (boundary predictions) for MODEL A and MODEL B, respectively. The star and triangle annotations represent labeling of MS data for training the models. The green curves represent predictions using MODEL B and the blue curve represents predictions using MODEL A.

[00023] FIGS. 8A, 8B and 8C show performance of MODEL A (blue, 2) and MODEL B (green, 3) against a known machine learning architecture trained on same data (rule-based method – red, 1). In Figure 10, each set of three bars is presented from left to right as: Rule-based, MAE 4.79s; Sequential NN, MAE 2,33s; and Ref-based NN, MAE 1.56s.

[00024] FIGS. 9A and 9B show additional sample labeled input and computed probabilities (boundary predictions) for MODEL A and MODEL B, respectively. The star and triangle annotations represent labeling of MS data for training the models. The green curves represent predictions using MODEL B and the blue curve represents predictions using MODEL A. Reference and sample peaks for a glycopeptide transition. **A**: In reference peak, grey line represents the input curve; red star and red triangle are human annotated peak start and end; blue/green solid line and dashed line indicate predictions from sequential/reference-based neural network model of peak start and end probabilities. **B**: Sample peaks for the same transition with different input curves. Note that reference-based models (green) output much more consistent predictions.

[00025] FIG. 10 relative standard deviations of boundary point selection for human annotators (gray, 1) versus MODEL A (blue, 2) and MODEL B (green, 3). Bar plots are of relative standard deviations (RSD) of 13 human annotators, and relative errors (RE) of sequential/reference-based neural network predictions. In Figure 10, each set of three bars is presented from left to right as: Sequential NN, Mean RE 0.034; Ref-based NN, Mean R=0.019; and Human Annotators, Mean = 0.025.

[00026] FIGS. 11A and 11B are scatter plots comparing abundance calculations between MODEL A and MODEL B, respectively. Scatter plot of peak abundance prediction/annotation from sequential (A) and reference-based(B) neural network. In sequential model predictions, weighted abundances (purple points) show higher consistency with ground truth and less outliers than argmax abundances (blue points), with a higher Pearson's r at 0.9969. Note that weighted abundances are biased upwards in low-abundance range. While in reference-based predictions results are highly similar.

[00027] FIG. 12 shows confidence scores for MODEL B and equation used to compute confidence score. Illustration of confidence score: blue dashed line (*) - model predicted peak

end probability; dark red line (#) - argmax prediction; light red lines (!) – other local maximums. Confidence score (0.634) is the area of shaded area divided by the area in the black box.

[00028] FIG. 13 shows for MODEL A histogram of prediction confidences, mean absolute errors of boundary predictions and scatter plots for different confidence score bins. **A**: Histogram of prediction confidences from sequential neural net-work. Abundance predictions of samples from four bins are illustrated in the following panels. **B**: Mean-absolute-error(MAE) of boundary predictions decreases steadily on samples with higher confidence. **C**, **D**, **E**, **F**: Scatter plots of peak abundance prediction/annotation in different confidence score bins. Samples with higher confidence achieved better correlation scores. Note that reported log-abundance Pearson's r values are calculated on the combination of samples from different transitions.

[00029] FIG. 14 shows for MODEL B histogram of prediction confidences and mean absolute errors of boundary predictions. **A**: Histogram of prediction confidences from reference-based neural network. **B**: Mean-absolute-error(MAE) of boundary predictions decreases steadily on samples with higher confidence.

DETAILED DESCRIPTION

[00030] In the description like reference numbers appearing in the drawings and description designate corresponding or like elements among the different views.

DEFINITIONS

[00031] As used in the present specification, the following words and phrases are intended to have the meanings as set forth below, except to the extent that the context in which they are used unambiguously indicates otherwise.

[00032] A "transition" refers to a temporal sequence of mass-to-charge ratio intensities from mass spectrometry. As used herein, the phrase a "transition" or a "multiple-reaction-monitoring (MRM) transition," refers to the mass to charge (m/z) peaks or signals observed when a glycopeptide, or a fragment thereof, is detected by MRM-MS. The MRM transition is detected as the transition of a precursor and product ion. The time interval for transition detection is dependent on the feed rate of the tandem liquid chromatography instrument and the flow rate within the MRM spectrometer.

[00033] A "biologic quantity" is understood as meaning an amount of an analyte in a biological sample or amount of a solute in a tandem chromatography instrument. Analytes

and solutes include, but are not limited to, glycans, peptides, glycopeptides, glycosylated peptides, proteins, glycoproteins, and fragments thereof.

[00034] As used herein, the phrase "biological sample," refers to a sample derived from, obtained by, generated from, provided from, take from, or removed from an organism; or from fluid or tissue from the organism. Biological samples include, but are not limited to synovial fluid, whole blood, blood serum, blood plasma, urine, sputum, tissue, saliva, tears, spinal fluid, tissue section(s) obtained by biopsy; cell(s) that are placed in or adapted to tissue culture; sweat, mucous, fecal material, gastric fluid, abdominal fluid, amniotic fluid, cyst fluid, peritoneal fluid, pancreatic juice, breast milk, lung lavage, marrow, gastric acid, bile, semen, pus, aqueous humor, transudate, and the like including derivatives, portions and combinations of the foregoing. In some examples, biological samples include, but are not limited, to blood and/or plasma. In some examples, biological samples include, but are not limited, to urine or stool. Biological samples include, but are not limited, to saliva. Biological samples include, but are not limited, to tissue dissections and tissue biopsies. Biological samples include, but are not limited, any derivative or fraction of the aforementioned biological samples.

[00035] As used herein, the term "glycan" refers to the carbohydrate residue of a glycoconjugate, such as the carbohydrate portion of a glycopeptide, glycoprotein, glycolipid or proteoglycan.

[00036] As used herein, the term "glycopeptide," refers to a peptide having at least one glycan residue bonded thereto.

[00037] As used herein, the phrase "glycosylated peptides," refers to a peptide bonded to a glycan residue.

[00038] As used herein, the phrase "glycopeptide fragment" or "glycosylated peptide fragment" refers to a glycosylated peptide (or glycopeptide) having an amino acid sequence that is the same as part (but not all) of the amino acid sequence of the glycosylated protein from which the glycosylated peptide is obtained by digestion, *e.g.*, with one or more protease(s) or by fragmentation, *e.g.*, ion fragmentation within a MRM-MS instrument. MRM refers to multiple-reaction-monitoring.

[00039] As used herein, the phrase "fragmenting a glycopeptide," refers to the ion fragmentation process which occurs in a MRM-MS instrument. Fragmenting may produce various fragments having the same mass but varying with respect to their charge.

[00040] As used herein, the phrase "multiple reaction monitoring mass spectrometry (MRM-MS)," refers to a highly sensitive and selective method for the targeted quantification

of glycans and peptides in biological samples. Unlike traditional mass spectrometry, MRM-MS is highly selective (targeted), allowing researchers to fine tune an instrument to specifically look for certain peptides fragments of interest. MRM allows for greater sensitivity, specificity, speed and quantitation of peptides fragments of interest, such as a potential biomarker. MRM-MS involves using one or more of a triple quadrupole (QQQ) mass spectrometer and a quadrupole time-of-flight (qTOF) mass spectrometer.

[00041] As used herein, "peptide," is meant to include glycopeptides unless stated otherwise.

[00042] "Retention time (RT)" is understood as meaning the time taken for a solute to pass through a chromatography column. It is calculated as the time from injection to detection. RT depends on several factors, including but not limited to, the gas flow rate; temperature differences in the oven and column; column degradation; column length; column type and carrier gas.

[00043] Mass Spectrometry (MS) data or datasets is understood as meaning a collection of spectra plotting intensity of transitions as a function of transition number. MS data for a given transition value is mass (m) over charge (z) intensity levels (m/z) over a time period (t), which is the retention time (RT).

[00044] An "abundance" is understood as meaning a computed amount of some biologic quantity represented in mass spectrometry data. An abundance may be computed by summing values at discrete time points between a start and stop peak boundary. Abundance may also be computed by integration over discrete time steps between the start and stop peak boundary. Integration may be performed by any standard method, e.g., $\Sigma\ (t_{i+1} - t_i) * (1/2) * [I(t_i) + I(t_{i+1})]$, i = 1.. n-1 (n is the number of discrete points between boundaries, $I(t_i)$ is mass-to-charge intensity at retention time t_i). Unless specified otherwise, abundance refers to relative abundance of one integrated MRM peak relative to another MRM peak.

[00045] The mathematical models that may be developed to practice aspects of the invention are limited to non-linear mathematical models in the sense that the non-linearity results in there being no set of optimized function / model parameters based on finding a global minimum value; rather, an optimized function / model parameters resides only in a local minimum (no global minimum exists). The model architectures contemplated include, but are not limited to, artificial neural networks, graph models and support vector machines.

[00046] A "machine learning model" or "learning model" is understood as meaning any variety of mathematical model having at least one non-linear activation layer (in addition to one or more affine transformations). A machine learning model is moreover trained or

optimized via minimization of one or more loss functions (e.g., minimization of cross-entropy loss or negative log-likelihood) that are separate from the model itself. A training or optimization process seeks to optimize the model to reproduce a known outcome (low bias) as well as enabling the model to make accurate predictions from unseen experiences (low variance). The model's output may be any variety of things relevant to the task, e.g., a probability or statistic, classification, sequence, etc. In one particular example, data used to train and optimize the model is sequential data such as time-series data where a target outcome is known, which sequential data with known target outputs is taken to mean labeled or annotated sequential data.

[00047] An example of a well-known model suited to make predictions based on labeled sequential data is a one-directional, or bi-directional recurrent neural network (RNN) with long short term memory (LSTM) units or gated recurrent units (GRUs), and optionally with an attention mechanism. An attention mechanism (or attention layer) may retain more important information from past state variables to influence a future state prediction than what may be possible in a final LSTM unit or GRU of a sequence, or a combination of LSTM units or GRUs along with a bi-directional RNN that represents the accumulated knowledge from past states. A attention mechanism applies learned weights to each of the input states to emphasize less or more important features of the input MS data points appearing in the sequence when generating output predictions. These learned weights are more or less directly coupled to each of the corresponding output states, or the final output state depending on the desired output.

[00048] Models that are trained and optimized using only labeled data are known as supervised models. Learning models may also be unsupervised or semi-supervised models. These models are trained and optimized using un-labeled data or a combination of labeled and unlabeled data. Models that are trained and optimized using combinations of labeled and unlabeled data or primarily un-labeled data are known as semi-supervised and unsupervised models, respectively. A learning model may be efficiently trained and optimized by the well-known approach of stochastic gradient descent (SGD). The learning model's hyper-parameters may be optimized by manual adjustment during training and/or validation, or learned by way of minimizing a surrogate loss function with respect to one or more of the hyperparameters, in parallel with the minimization of the learning model's global loss function.

[00049] A "training set" is a dataset used to train or tune parameters in a learning model to reproduce a known, targeted output.

[00050] A "validation set" is a dataset used to tune the model's predictive capabilities against new experiences to arrive at an optimal balance between matching the labeled data (low bias) and predicting start/stop locations for unseen transitions (low variance). The validation set is applied after the model has been tuned based on the training set.

[00051] A "test set" (or leave-out validation set) is a dataset used to evaluate the optimized model's predictive capabilities. The model's performance against the test set is the metric used to assess the model's ability to make an accurate prediction for new, unseen transitions.

[00052] "Featurization" is understood as meaning processes associated with converting raw data representing samples from a distribution for use as input to a machine learning model. Examples of featurization include discretization of an analog signal into a digital signal, filtering or de-noising or clipping parts of data (e.g., reducing signal-to-noise ratio), normalizing signals and standardization of signals (e.g., centering a target peak within a window). Featurization may also include introducing noise into a data set for purposes of validating a model or related augmentation of samples for generalization and/or improving a model's ability to make predictions from new (or unseen) experiences.

[00053] According to the disclosure the dataset of sequential data from which training, validation and test datasets were sampled may include at least labeled start and stop locations for transitions, each of which locations indicate boundary points for computing an abundance by integration between the boundary points. In a preferred embodiment the targeted output are probabilities of start and stop locations for each point in the sequential data of a transition.

DISCUSSION

[00054] As mentioned earlier, traditional analysis of MRM experiments involves choosing a start and stop of a chromatographic peak by hand and working from transitions previously characterized by data-dependent acquisition or other discovery techniques. Software suites, such as Agilent's MassHunter software, plot signal intensities over a pre-specified range of retention times (RT) for a given precursor and product mass-to-charge ratio. Manual selection of the beginning and end of the peak yields an observed RT, peak width, and integrated abundance value. Person to person variability in assessment, human error, and a large time investment however render this method inadequate for non-research use.

[00055] For transitions characterized by clearly defined single peaks and high signal-to-noise ratio, the above manual process of picking boundary points for computing

abundances may be effective, but this method may not provide accurate or produce consistent results when detecting abundances for biologic quantities characterized by a low signal-to-noise ratio.

[00056] Referring to FIG. 1, according to the disclosure a trained machine learning model is instead used to automatically select boundary points in mass spectra for a specified retention time window and mass/charge intensity. The input to the machine learning model is a sequence of points (sample of MS data). Each point represents a mass/charge intensity for a corresponding retention time.

[00057] Referring to FIG. 2, according to one implementation the machine learning model is a neural network that produces probabilities of peak start and peak stop times for each of the corresponding points provided as input. The neural network is trained using mass spectra that has target peak start and stop retention times labeled. Through a training and validation process the optimal model parameters are found. The labeled data may be obtained from human annotators selecting peak start / stop times using a graphical user interface and input device.

[00058] Referring to FIG. 3 there is a process flow for featurizing raw mass spectra data for input to the machine learning model of FIG. 1. The featurization steps may include (1) remove mass / charge intensities above a certain threshold amount or, alternatively, choosing only those points that are mostly noise and near a signal-to-noise transition region of the MS data; and (2) discretizing the analog signal into a set of points to produce sequential data.

[00059] Referring to FIG. 4, in one particular implementation the model computes an abundance amount based on a weighted sum of integrals. One of the most direct ways is to interpret the machine learning model's prediction outcomes as probability distributions. In other words, to have the model evaluate probabilities of all possible pairs of points in a sequence that could serve as peak start/end markers. For a sequence of N points, this involves in total N x (N-1)/2 pairs after imposing the order constraint. Each pair's probability can be easily calculated by multiplying prediction values of the peak start/end tasks on the two points. The probabilities are then used to calculate a weighted sum over the integrals between corresponding pairs, providing a mean estimate instead of a single abundance value calculated from the two end points having the highest probability of being the true boundary. A mean of the abundances may provide more robust abundance estimates.

Embodiments

[00060] In some examples, set forth herein is a method of training a model to identify mass spectrometry peaks, comprising: using a plurality of labeled sequential data, each of the plurality of labeled sequential data comprising a peak start and a peak stop label representing retention time values characteristic of glycopeptides or peptides, or fragments thereof, and training at least one machine learning model using the labeled sequential data, wherein the trained model is adapted for identifying mass spectrometry start and stop peaks indicating the presence of one or more glycopeptides or peptides, or a fragment thereof, present in a biologic sample.

[00061] In some examples, including any of the foregoing, the method further comprises quantifying the presence of the one or more glycopeptides or peptides by integration over time between the mass spectrometry start and stop peaks from the biologic sample.

[00062] In some examples, including any of the foregoing, the machine learning model is a recurrent neural network model.

[00063] In some examples, including any of the foregoing, the recurrent neural network is a bi-directional LSTM recurrent neural network model with attention layer.

[00064] In some examples, set forth herein is a method comprising computing an abundance for a glycopeptide or peptide using a machine learning model trained a process set forth herein.

[00065] In some examples, set forth herein is a system, comprising a data repository of labeled sequential data; one or more processors coupled to the repository; machine executable code, residing in non-transitory memory accessible to the one or more processors, that when executed by the one or more processors performs a method set forth herein.

[00066] In some examples, set forth herein is a method of identifying mass spectrometry peaks, comprising: providing at least one trained neural network model that produces an output as a function of input parameters, wherein values for the input parameters comprise a sequence of retention time values for a glycopeptide having a m/z (mass to charge) ratio; and providing as values for the input parameters mass spectrometry (MS) data comprising mass-to-charge (m/z) ratios over the sequence of retention time values from a mass spectrometry of a biologic sample; wherein the output comprises a location for one or more peak start and peak stop retention time values indicating the presence of one or more analytes, or a fragment thereof, in the biologic sample.

[00067] In some examples, including any of the foregoing, the method further

comprises quantifying the presence of the one or more glycopeptides or peptides by summing or integration over retention time values between the mass spectrometry start and stop peaks from the biologic sample.

[00068] In some examples, including any of the foregoing, the method further comprises quantifying the presence of the one or more glycopeptides or peptides by summing over retention time values between the mass spectrometry start and stop peaks from the biologic sample. In some examples, including any of the foregoing, the method further comprises quantifying the presence of the one or more glycopeptides or peptides by integration over retention time values between the mass spectrometry start and stop peaks from the biologic sample.

[00069] In some examples, including any of the foregoing, the quantifying step includes applying a weighted smoothing function over the predicted start and stop peaks.

[00070] In some examples, set forth herein is a system, comprising a data repository of mass spectrometry (MS) data; one or more processors coupled to the repository; machine executable code, residing in non-transitory memory accessible to the one or more processors, that when executed by the one or more processors performs a method set forth herein.

[00071] In some examples, set forth herein is a method comprising: featurizing data from raw MS data, wherein the raw MS data comprises a plurality of analog samples, each representing mass/charge intensity, the featurizing comprising, for each of the analog samples, the steps of centering the sample within a retention time window, discretizing the sample into a sequence of points representing intensity values, standardizing the intensity values, and assigning labels to points among the sequence of points corresponding to peak start and peak stop times to produce a labeled sequence of points; wherein the labeled sequence of points are configured for training a machine learning model to predict an abundance in unseen MS data.

[00072] In some examples, including any of the foregoing, the centering comprises positioning an apex of a curve at a center of the retention window.

[00073] In some examples, including any of the foregoing, the standardizing the intensity values excludes intensity values greater than a threshold amount. In some examples, including any of the foregoing, the threshold amount is about 500.

[00074] In some examples, set forth herein is a system, comprising a data repository of raw mass spectrometry (MS) data comprising a plurality of analog samples; one or more processors coupled to the repository; machine executable code, residing in non-transitory memory accessible to the one or more processors, that when executed by the one or more

processors performs the method of any In some examples, set forth herein is a method, comprising computing a distribution of probabilities over a sequence of points representing mass spectrometry (MS) data for a given m/z ratio and retention time period, the probabilities representing a likely peak start or peak stop retention time, and computing a statistic for the abundance using the distribution of probabilities, wherein the statistic is indicative of the presence of a biologic sample in the MS data.

[00075] In some examples, including any of the foregoing, the statistic is a mean abundance computed as a weighted sum of integrals using all possible pairs of start and end points for evaluating the integral.

[00076] In some examples, including any of the foregoing, the machine learning model is a sequential model (MODEL A) or a reference-based model (MODEL B).

[00077] In some examples, set forth herein is a method, comprising: using a plurality of labeled sequential data, and training at least one machine learning model using the labeled sequential data, wherein the trained model is adapted for identifying a time period or frequency range where a start and stop peak of an intensity occur in an unseen sample.

[00078] In some examples, including any of the foregoing, the start and stop peak is contained in a noise to signal transition time period.

[00079] In some examples, including any of the foregoing, the machine learning model is a recurrent neural network with an attention layer.

[00080] In some examples, including any of the foregoing, the sample represents a physical thing such that a summation or integration over time or frequency from the start peak to the stop peak of the intensity quantifies the physical thing.

[00081] In some examples, including any of the foregoing, the sample is a sequential series of mass to charge intensities over a specified time period and the summation or integration represents the abundance of the physical thing.

[00082] In some examples, set forth herein is a method, comprising: providing at least one trained neural network model that produces an output as a function of input parameters, wherein values for the input parameters are a temporal or spectral sequence of intensity values representing a quantity; and providing values for each of the input parameters, the values representing an unseen sequence of intensities for a quantity; wherein the output comprises one or more peak start and peak stop times or frequencies, which are used to quantify the quantity.

[00083] In some examples, including any of the foregoing, the quantity is quantified by summing or integration over time or frequency between the peak start and peak stop times or

frequencies, respectively.

[00084] In some examples, including any of the foregoing, the intensities are obtained from a process selected from the set of mass spectrometry, MRM, liquid or gas chromatography, x-ray diffraction, Raman spectroscopy, UV-VIS spectroscopy, fluorimetry spectroscopy or phosphorimetry spectroscopy, Nuclear Magnetic Resonance (NMR) spectroscopy, or Magnetic Resonance Imaging (MRI).

EXAMPLES

[00085] The following example describes a study conducted in accordance with several aspects of the disclosure.

[00086] In this study two learning models were trained, validated, and then tested to assess accuracy in predicting the location of peaks (or more precisely, the start/stop boundaries of peaks in order to compute more accurately an abundance for the associated transition). The particular architecture chosen for the learning model was a bi-directional recurrent neural network with long short term memory units and attention layer (RNN-LSTM-AL). The model architecture and training method is illustrated in FIGS 5A, 5B, 6A, 6B and 6C. The model's input parameters or input state variables correspond to a temporal sequence representing mass-to-charge ratio (m/z) intensities.

[00087] The study accessed a large dataset of manually-labeled peaks from Mass Spectrometry (MS) transitions spanning a wide dynamic range, including both peptides and intact glycopeptides. This dataset contained the temporal sequence of intensities within each transition's Retention Time (RT). Each one of these transitions is a sequential data input to the RNN-LSTM-AL, and the associated human-annotated (or labeled) start and stop RT (indicating boundary points for integration to derive abundance) included with each one of the training, validation, and test transitions are the labels used for supervised learning, validation and testing, respectively.

2.1 Data Sets

[00088] Two mass spectrometry datasets were used for the study. Raw Agilent QQQ MS output for the first experiment consisted of 716 transitions assessed in 210 human serum samples, which yielded 106355 valid sample peaks after filtering for positive abundance and RT shift. This set was split into a training set with 572 transitions and a validation set with 144 peaks.

[00089] The second experiment used MS output generated from 135 human serum samples purchased from a commercial biorepository. 503 transitions were evaluated, yielding

67672 high quality peaks that were used exclusively for testing. 313 test set transitions were not present in the train/validation set. The 313 test set transitions were used to compare model output predictions.

2.2 Baseline Method

[00090] The performance of the trained and validated RNN-LSTM-AL models were compared to two baseline predictors. The first of these baseline predictors is Skyline, which is a well- known open-source software tool used for data analysis of mass spectrometry applications including multiple reaction monitoring. The Skyline software calculates peak areas (abundances) of test samples. The second baseline predictor for comparison with the RNN models is a deterministic method of prediction and will hereinafter be referred to as the "rule-based"' method. The architecture for the rule-based method is described in Ken Aoshima, Kentaro Takahashi, Masayuki Ikawa, Takayuki Kimura, Mitsuru Fukuda, Satoshi Tanaka, Howell E Parry, Yuichiro Fujita, Akiyasu C Yoshizawa, Shin-ichi Utsunomiya, et al. *A simple peak detection and label free quantitation algorithm for chromatography-mass spectrometry*. BMC bioinformatics, 15(1):376, 2014. The rule-based method model was developed for this study by tuning its model parameters to maximize train and validation set performances. And these tuned parameters were then applied during test set evaluations.

2.3 Sequential models for Peak Integration

[00091] The RNN-LSTM-AL architecture was selected for making predictions from sequences of MS data in the study.

[00092] A visual and block description of the training / prediction process and flow for the first of the two RNN-LSTM-AL models used in the study (hereinafter "MODEL A" or "sequential model") is illustrated in FIGS. 5A and 5B. The inputs (mass/charge ratio intensity as a function of time) are encoded through a bi-LSTM layer and attention mechanism for relating separate positions. The prediction is performed point-wise, meaning for every input the model predicts an output. The model was trained to generate a sequence of outputs with dimensionality of N x 2, where "N" is the number of input parameters (points corresponding to the retention times where an intensity value was input to the model, ranging from 50 to 300 depending on the nature of the curve in the sample, see Sec. 5.3, *infra*) for MODEL A, the number of data points in the sequence, and output values corresponding to each input value. The two (2) values output for each of the N points are probabilities (ranging from 0 to 1) that the corresponding point represents a peak start or a peak stop. A point, or neighborhood of points output from the model with values closest to 1 (high probability) for a given transition

are then used to define a boundary for the interval over which the abundance integral (integrated over time) is computed.

[00093] The output probabilities from the model thus form distributions of boundaries. The model is trained end-to-end by gradient descent on the cross entropy loss between the predicted boundary and the ground truth, or manually-labeled peak start and speak stops. Minimum hyper-parameter search, e.g., Bayesian optimization applied to an objective function minimized with respect to the model's hyper-parameters, is applied to optimize performance on the leave-out validation or test set, in order to achieve good generalization and avoid overfitting to the training set. The featurization and hyper-parameters are discussed in more detail below.

2.4 Reference-Based Sequential Models

[00094] In some applications, due to the highly variable shape and width of transition peaks, a unified prediction model such as MODEL A may not generalize well to unseen transitions. Based on the observation that samples from different patients on the same transition are consistent in shape, a second RNN-LSTM-AL model (hereinafter "MODEL B" or "reference-based model") was also developed.

[00095] A visual and block description of the training / prediction process and flow for MODEL B is illustrated in FIGS. 6A, 6B and 6C. MODEL B takes a query sample and reference as inputs, respectively. The encoded features for both are compared and merged in the sample-reference attention layer, generating predictions. Provided with query peak features, reference peak features and reference labels, the attention layer will compare query and reference point by point such that points in the query that are highly similar to the labeled start/end points in the reference will obtain higher predicted probabilities. Given two consistent samples, the ideal attention map of this operation will be an "identity matrix," in which points with the same surroundings are regarded as highly similar (see FIG. 6C). A divergence term between the attention map and a retention time mapping matrix was included in the optimization process to achieve good regularization.

[00096] Note that in the setting of reference-based prediction (as opposed to the query-based prediction), the task is no longer trying to identify the noise-signal transition point as with MODEL A, but rather a good encoding mechanism of a point in the sequence and its context in order to optimize the similarity mapping. In applications, reference-based models act similarly to few-shot learning models. They are trained on pairs of query/reference, aiming at accurately mapping reference labels to predictions on the query sample. This

naturally allows the model to perform better on unseen transition peaks as long as a single (or a few) correctly labeled references are provided.

[00097] The adopted structure for MODEL B may be most applicable and cost-efficient for high-throughput experiments on the same set of transitions, in which manual annotations of a few high quality reference peaks are a reasonable prerequisite. On most discovery or small scale experiments a standalone reference-free predictors may be preferred.

[00098] Referring to FIGS. 7A and 7B, each figure has two plots. The first plot is a plot of a transition with labeled start and stop times indicated by the star and triangle, respectively. The bottom plot is a plot of the probabilities of start / stop times predicted by the trained and validated models.

3.1 Boundary and Abundance Prediction performance

[00099] Leave-out or test set samples were used for evaluating the models. Several metrics were used to present a comprehensive view of the problem being solved. Predictions are evaluated on two criteria: (1) accuracy in predicting boundaries (peak start/end) and (2) accuracy in predicting abundances (peak area). While the former is directly related to the setup of the problem, the latter will be of much higher importance to downstream applications which utilize analyte abundance for analysis.

[000100] In predictions, points in the curve with maximum predicted probabilities for the two tasks, i.e., the tasks of finding a peak start time and a peak end time, are used as markers for boundaries, which will be referred to as "argmax boundaries" in the following text. We calculated mean-absolute-error (MAE) between argmax boundaries and the human annotations or labels, along with an accuracy score defined as the ratio of samples whose boundary predictions are within an error threshold of 1.2 seconds (2 points in a sequence) around the ground truth annotations (human labeled points). Given that the peak duration is 20 +/- 8 seconds (33 +/- 14 points in a sequence) in the test set, the error threshold is small enough so that any boundary prediction within the window will not significantly change the abundance.

[000101] For evaluation of abundances, first calculated are baseline-adjusted integrals between boundaries from ground truths (human labeled) and model predictions respectively, and then evaluated Spearman's rank correlation coefficient (Spearman's r) and Pearson correlation coefficient (Pearson's r) between the two on each individual transition. Due to the wide range of abundances ($1 \sim 10^{15}$), Pearson's r will be dominated by samples with large abundances, here reported are correlations calculated on log-abundances instead to avoid

bias. Average scores over all transitions are reported. Furthermore, for indication of practical usage, also reported is an accuracy score defined as ratio of samples whose predicted abundances are within +/- 5% of the annotated abundances. Note that this accuracy metric has one caveat : integrals computed over an interval defined by boundary predictions that deviate from the annotated boundaries may appear to have very similar abundances as the annotated abundances because start/stop boundary prediction errors may cancel each other out. That is, the % error in the predicted peak start time may cancel out a corresponding % error in the predicted peak stop time when the intensity values are summed in the integral.

TABLE 1: Prediction performance on leave-out test set (67672 peaks)					
	Model Boundary Prediction		Abundance Prediction		
	MAE (second)	Accuracy*	Log-Abundance Pearson's r	Spearman's r	Accuracy**
Skyline	N/A	N/A	0.6388	0.7372	0.4365
Rule-based	4.79	0.285	0.9879	0.9708	0.5847
MODEL A	2.33	0.432	0.9897	0.9835	0.6889
MODEL B	1.56	0.584	0.9974	0.9917	0.7995

Table 2: Prediction performances on train set (86311 samples)					
Model	Boundary Prediction		Abundance Prediction		
	MAE (second)	Accuracy*	Log-Abd Pearson's r	Spearman's Rho	Accuracy**
Rule-based	3.97	0.311	0.9097	0.9924	0.645
MODEL A	1.79	0.601	0.9477	0.9965	0.743
MODEL B	1.54	0.667	0.9592	0.9969	0.779

* Ratio of samples whose boundary predictions are within 1.2 seconds error

** Ratio of samples whose abundance predictions are within 5% error

Table 3: Prediction Performances on validation set (20044 samples)					
Model	Boundary Prediction		Abundance Prediction		
	MAE(second)	Accuracy	Log-Abd	Spearman's	Accuracy

			Pearson's r	Rho	
Rule-based	4.08	0.320	0.9141	0.9023	0.6593
MODEL A	2.57	0.445	0.9332	0.9270	0.6884
MODEL B	2.80	0.372	0.9541	0.9398	0.6563

[000102] TABLE 1 presents performance scores of the three tested methods and Skyline predictions. Note that Skyline generated abundance predictions only and was not evaluated and compared on boundary prediction tasks. Both MODEL A and MODEL B demonstrated significant performance boosts on all tasks over the rule-based method and Skyline predictions. MODEL B, with the aid of extra reference data, achieved top scores. On boundary predictions, as the rule-based method is tuned to optimize training/validation set performances (see TABLE 3), strong overfitting is observed. In comparison, thanks to the better sample quality, MODELS A and B achieved MAE of 2.33 and 1.56 seconds on the test set, respectively (with and without reference), two times smaller than the rule-based method. Likewise, in accuracy score performance MODEL B took the lead with a 30 % advantage over the rule-based method. FIGS. 8A and 8B show a bar plot and un-normalized cumulative distribution of boundary prediction error, respectively, in which MODELS A and B presented distributions with lower error as well as less outliers.

[000103] Abundance predictions presented similar results: MODELS A and B achieved better and more robust performances on evaluating peak abundances (FIG. 8C), with an over 20% increase on accuracy. Note that in this problem influences from errors in boundary prediction will be largely alleviated as boundary points typically contribute little to the overall integrals. Correlation scores indicated that MODELS A and B performed significantly better than current off-the-shelf mass spec data analysis software, with a near-perfect Spearman's correlation coefficient at 0.9975 and 0.9990. Intriguingly, the rule-based method also achieved much higher correlation score than the Skyline predictions. Without wishing to be tied to any theory, this result may have been due to the complex composition of test samples, which generated co-eluting peaks or compounds with close retention time. When the RT window was set closely around the desired transition (which is also a setting typically used by human annotators), Skyline might have accidentally pick peaks from different range, resulting in significantly worse performance.

[000104] Comparing MODEL A with MODEL B, because the extra “reference data” was included during the training and validation phases for MODEL B, this model performed

better than MODEL A, offering 15% and 10% increases on the accuracy of boundary and abundance predictions. MODEL B presented better boundary error distribution with fewer outliers (2.2% versus 7.1%, FIG. 8B), consistent with the observation in abundance correlation plots: green points are more concentrated around the center line while blue points include a few outlying predictions (FI9G. 8C).

[000105] Depending on the application however MODEL A may be preferred, since less training sets are needed. The additional reference data used with MODEL B may not be needed, or practical for small scale applications or experiments since MODEL A may provide good accuracy / predicted abundance as the results show. MODEL B may be preferred for larger scale experiments or when more accuracy is needed.

[000106] In these experiments, MODEL B provided another advantage not displayed in the plots: higher consistency in predictions on the same transitions across multiple samples. Especially for transitions whose boundaries could vary due to shoulder and valley points, the MODEL B tends to follow the same prediction form of its reference (FIG. 9A), which is the main objective of its model structure design. This is highly desirable in high throughput experiments in that abundances are generated based on the very same standard and hence allowing for more accurate comparison and quantification. But it should also be noted that the strong prior assumption in the model may also hurt its performance in cases of retention time shift or strong batch effect.

3.2 Comparison with human annotators

[000107] To further validate the practical usage of neural network-based models, we compared model predictions against a group of human annotators calculating peak abundances.

[000108] Within the test set, analyzed was a subset consisting of 12 transitions for all 135 serum samples, and 12 annotators marked the peak start/end independently. Combined with the original test set labels, 13 sets of human annotations are prepared for this subset. Then we calculated relative standard deviations for each sample peak across the 13 annotations as a proxy for variations between different annotators, and compared with MODEL A's and MODEL B's predictions relative errors.

[000109] Results are summarized in FIG. 10. High consistency over annotators is observed over most samples, as demonstrated by the grey bars FIG. 10. Mean RSD over all sample peaks for human annotators is 2.5 %. In parallel, both MODEL A and MODEL B performed well on the subset, achieving mean relative error of 3.5 % and 2.1 % respectively

compared with human annotator average. The distributions of error, as shown in the blue and green bars, are similar but slightly worse than the RSD across annotators. Wilcoxon signed-rank test doesn't show significance ($p=0.11, 0.50$) for both models, indicating that the differences are not significant.

[000110] Viewing predictions from each individual annotator as a group, we then calculated pairwise differences between annotators: average of the relative abundance differences on each sample peak. The outcomes show a discrepancy between annotators at 0.025 +/- 0.009, in line with the RSD value. At the same time, difference between the MODEL B prediction and the 13 annotators is 0.026 +/- 0.007, overlapping with the inter-annotator difference. Corresponding value for MODEL A is 0.039 +/- 0.005. Given that the model vs. human difference is not significantly larger, or even close in the reference-based case, it is clear that the neural network-based models can serve as good substitutes for manual annotations in this task. We also note that the human annotators had a peak with selected start and stop to serve as a reference for their reads, making their task directly analogous to the MODEL B.

3.3 Weighted Peak Start/End for Abundance Calculation}

[000111] In peaks with shoulder and valley points, determination of boundaries will suffer from a much higher variability. Even in human annotations variations across different annotators exist, so it is expected that output from MODEL A and MODEL B will present multimodal shapes due to the averaging over the ensemble of training samples (FIG. 9B, FIG. 7B). This is rather severe in reference-free cases (MODEL A) as there doesn't exist a strong reference prior, causing results to sometimes be arbitrary when applying the argmax boundaries. In these cases methods were reviewed that can better accommodate full predictions instead of a single point estimate of boundaries.

[000112] One of the most direct ways is to interpret the prediction outcomes as probability distributions. In other words, to have the model evaluate probabilities of all possible pairs of points in a sequence that could serve as peak start/end markers. For a sequence of N points, this involves in total N x (N-1)/2 pairs after imposing the order constraint. Each pair's probability can be easily calculated by multiplying the prediction values of the peak start/end tasks on the two points. The probabilities are then used to calculate the weighted sum over the integrals between corresponding pairs, providing a mean estimate instead of a MAP estimate (this alternative method for computing abundances will hereinafter be referred to as the “weighted sum method”). A MAP estimate computes one

abundance value using the argmax values as endpoints.

[000113] A mean of the abundances may provide more robust abundance estimates than using argmax boundaries, especially when a model architecture more similar to MODEL A than MODEL B is used. As no explicit boundaries are used, the method may avoid issue of peak end prediction appearing prior (or very close) to a peak start prediction. A comparison of abundance correlation for MODEL A and MODEL B is presented in FIGS. 11A and 11B, respectively.

[000114] Especially for MODEL A, the weighted sum method helped in generating more accurate abundances, raising abundance correlation from 0.9897 to 0.9969. Increase in MODEL B is relatively minor (from 0.9974 to 0.9977). The post-processing step mainly helped in avoiding outliers whose argmax boundaries caused significantly lower abundances than human annotations. These samples typically suffer from abnormally high predicted boundary probabilities within the range of peak (at valley or shoulder points), which are then chosen by the argmax operation. At the same time, the weighted sum method caused overall predictions to bias upwards for low-abundance peaks, which is mainly due to noisy signals and predictions. In cases where signal-noise ratio is low, predicted boundary probability could fluctuate on a longer retention time range at the periphery of a peak (FIGS. 7A and 7B, and FIG. 12), leading to a higher weighted sum. We deemed this effect to be acceptable as noisy peaks will be of less value for downstream applications.

[000115] From these rules, confidence values were calculated as follows: for prediction on a sample curve with N points, calculate all local maximums in the range. The global maximum is labeled as

$$(\hat{x}^*, \hat{y}^*), \text{ in which } \hat{x}^* \in \{1, 2, ..., N\}, \hat{y}^* \in [0, 1]$$

[000116] Local maximums (if they exist) on its left are labeled as $(\hat{x}_i, \hat{y}_i), i = 1, 2, ..., p,$ in which $\hat{x}_1 < \hat{x}_2 < ... < \hat{x}_p,$ local maximums on the right (if they exist) labeled as $(\tilde{x}_j, \tilde{y}_j), j = 1, 2, ..., q,$ in which $\tilde{x}_1 > \tilde{x}_2 > ... > \tilde{x}_q$. By further defining $\hat{x}_0 = \hat{x}^*$ and $\tilde{x}_0 = \hat{x}^*$ the confidence is calculated as:

$$Confidence = \hat{y}^* - \frac{1}{N-1}\left(\sum_{i=1}^{p}(\hat{x}_i - \hat{x}_{i-1}) \times \max(\hat{y}_i, \hat{y}_{i+1}, \ldots, \hat{y}_p) + \sum_{j=1}^{q}(\tilde{x}_{j-1} - \tilde{x}_j) \times \max(\tilde{y}_j, \tilde{y}_{j+1}, \ldots, \tilde{y}_q)\right)$$

.

A graphic illustration is shown in FIG. 12.

3.4 Prediction Confidence

[000117] Another core advantage of the point-wise prediction structure is ease in estimating prediction certainty. Though internal ways of quantifying uncertainty in neural network predictions are still complex and come with restrictions, it is fairly straightforward in this task to infer certain/uncertain by inspecting model predictions for boundaries. Most predicted probabilities will take the form of Gaussian mixtures, the quantity and heights of which represent the model's confidence. In the study a heuristic method was adopted to quantify prediction confidences based on the guidelines below:

item Lower predicted values -> lower confidence;

item Multiple local maximums -> lower confidence;

item More dispersed local maximums -> lower confidences.

[000118] From these rules, confidence values were computed. FIG. 12 provides a description of the math used to compute the confidence level, along with a graphical illustration. The confidence score is proportional to the shaded area. One wants the prediction values to be high (close to 1) and unimodal, or close if multiple local maximums appear.

[000119] Confidence scores were computed for all predictions for MODEL A and MODEL B (FIG. 13A and 14A). Across different confidence cutoffs, one can see a trend of performance change. As shown in FIGS. 13B and 14B prediction errors on boundaries steadily decrease with increasing confidence, dropping to near 1 second on the most confident samples.

[000120] Correlation of abundances on different confidence intervals are illustrated in FIGS. 13A – 13F. Predictions on high-abundance samples typically have higher confidence and align better with human annotations. The top two bins predictions achieved Pearson's r over 0.999. In contrast, samples with lower prediction confidence yield worse performance,

with higher boundary mean absolute errors (MAEs) and lower abundance correlations. This is in part due to the worse signal-noise ratio of input samples as indicated by the low abundances. Overall, the measurement of confidence served as a good indicator of noise scale and model performance.

[000121] This post-processing step may provide versatility in implementing workflow quality control. In real world applications, a direct usage would be setting up a threshold confidence score such that predictions below the value should be excluded or presented to human experts for reevaluation. This would also be helpful for detecting abnormalities in experimental data.

5.0 Experimental

Experimental Parameters

[000122] Chemicals and Reagents: Human serum, dithiothreitol (DTT) and iodoacetamide (IAA) were purchased from Sigma-Aldrich (St. Louis, MO). Sequencing grade trypsin was purchased from Promega (Madison, WI). Acetonitrile (LC-MS grade) was purchased from Honeywell (Muskegon, MI).

Sample Preparation: Serum samples were reduced with DTT and alkylated with IAA followed by digestion with trypsin in a water bath at 37C for 18 hours. To quench the digestion, formic acid was added to each sample after incubation to a final concentration of 1 % (v/v).

5.1 LC-MS Analysis

[000123] Digested serum samples were injected into an Agilent 6490 triple quadrupole mass spectrometer equipped with an Agilent 1290 infinity UHPLC system and an Agilent ZORBAX Eclipse Plus C18 column (2.1 mm x 150 mm i.d., 1.8 um particle size). Separation of the peptides and glycopeptides was performed using a 70-min binary gradient. The aqueous mobile phase A was 3 % acetonitrile, 0.1 % formic acid in water (v/v), and the organic mobile phase B was 90 % acetonitrile 0.1 % formic acid in water (v/v). The flow rate was set at 0.5 mL/min. Electrospray ionization (ESI) was used as the ionization source and was operated in positive ion mode. The triple quadrupole was operated in dynamic multiple reaction monitoring (dMRM) mode. The peptide and glycopeptide transitions employed in training and testing were a selection of those published in in Li, *et al.*, Site-Specific Glycosylation Quantification of 50 serum Glycoproteins Enhanced by Predictive Glycopeptidomics for Improved Disease Biomarker Discovery, *Anal. Chem.* 2019, 91, 5433-

5445; DOI: 10.1021/acs.analchem.9b00776, the entire contents of which are herein incorporated by reference in its entirety for all purposes.

5.2 Dataset Preparation

[000124] Training and validation set peaks and test set peaks were collected from two sets of mass spectra experiments. Set one consisted of human serum samples and transitions measured in Carlito Lebrilla's lab at UC Davis in late 2017, while set two were commercially-available serum from ovarian cancer and benign mass patients purchased from Indivumed GmbH in mid-2018. Both sets of samples underwent the experimental protocol outlined above. Post-run, Agilent.D format files were converted to mzML via msConvert within the Proteowizard 3.0 software suite for bioinformatic processing.

[000125] To obtain human annotations, all transitions were labeled by a mass spectra specialist for one reference serum sample (pooled Sigma serum), and these labels make up the "reference" peaks. For the remaining samples, an annotator labeled the peak start and peak end retention time marks in Agilent MassHunter Quantitative Analysis B.5.0, based on the reference peak and accounting for any shift in retention time.

[000126] From the two experiments, only peaks labelled with positive abundances were selected. Samples with large retention time shift (> 0.2 min) were also excluded. In total 106355 (train/validation) and 67672 (test) peaks passed sanity tests and were used in this study. For each transition, we collected all signals within a retention time window from reference retention time start - 0.2 minutes to reference retention time stop + 0.2 minutes, allowing for a mass-charge ratio window of +/- 0.1 around the desired precursor and product m/z. Signals were summed along the mass-charge ratio window, producing an extracted-ion chromatogram (XIC). Both sequential models and human annotators are presented with the XIC curves cropped around the target transition (FIGS. 7A, 7B and 9B show some example XIC inputs).

[000127] The test set used in the above section entitled “Comparison with human annotators” contained 1619 peaks. All peaks were further annotated by 12 annotators independently for consistency, but only annotations with positive abundance were used to calculate transition-wise relative standard deviation and pair-wise discrepancies between annotators.

5.3 Featurization

[000128] Inputs are formulated as sequences of points on the intensity-retention time plane, with no fixed start/stop or length. To generate a uniform representation, featurization

was applied on each sample curve. Retention times for all points in a single sample were first centered at the sample curve's apex position, then expanded by 128 equally spaced Gaussian bins, ranging from -1 min to 1 min, generating a vector F_{RT} of 256 real numbers representing time points.

[000129] Mass to charge intensity values have a wide range, intensity values were similarly discretized by applying 256 Gaussian functions on the range from 0 to 500, turning the analog values into a vector I_{RT} of 256 real values representing intensity. Adopting the assumption that any points with intensities larger than the top threshold should be regarded as part of a peak, only those points located in noise or noise-signal transition stages were featurized. The full input X for each sample curve (i.e., input parameters or states to the RNN model) after forming F_{RT} and I_{RT} was formed is matrix of size N x 384. The variable N ranged from 50 to 300, representing the length of the curve.

[000130] For MODEL B, each sample is paired with its reference. The pair of inputs will have the same featurization process as above and were formed as two matrices X and X_{ref} of dimensions N x 384 and N' x 384. N and N' should be close but they are not necessarily equal to each other.

[000131] Labeling of samples were formulated as the same sequential structure, in which human annotations of peak start/end will have values of 1 on corresponding tasks. Further label smoothing was applied to increase stability and robustness: sample points close to the human annotated label / ground truth start/end annotations will also have positive label values, calculated as:

$$\tilde{y} = \exp\left(-\frac{d_{RT}^2}{\sigma_{RT}} - \frac{d_{ABD}^2}{\sigma_{ABD}}\right)$$

d_{RT} and d_{ABD} are differences in retention time (RT) and differences in abundance (ABD) and σ_{RT} and σ_{ABD} are hyperparameters controlling the respective decay rates during training. The purpose of smoothing is to allow model adaptivity to noisy labels and/or account for human labeling / annotations that do not specify a single point. A neighborhood of points near the label however can represent reasonable points indicating transition from noise to signal as well.

5.4 Training Settings

[000132] Referring again to FIGS. 5A and 5B (MODEL A) and FIGS. 6A and 6B (MODEL B), MODEL A and MODEL B were implemented in Pytorch.

[000133] MODEL A contained two bi-LSTM layers with 128 features for hidden state, and LSTM outputs are then applied to a multi-head attention layer with 8 heads and 32

features in each head. Outputs from heads are concatenated and go through a fully-connected hidden layer with 32 units, then map to 2 tasks each with two classes through an affine layer. Softmax is applied to each task and each point in the sequence for generating probabilities of peak start/end. The model is trained by cross entropy loss between smoothed labels and the softmax predictions on the train subset containing 86311 samples for 20 epochs, with batch size of 8 and learning rate of 0.001. Early stopping is applied based on validation set performances.

[000134] MODEL B contained two sets of sequential network structure identical to the above, except for the fully connected layer, used to encode sample and reference respectively. Weights are not shared between the two structures. A sample-reference attention layer is applied on the sample encodings (as queries), reference encodings (as keys) and reference labels (as values), the output of which will be of the same length as sample encodings and 2 on the task dimension. This is directly used as model predictions for peak start/end probabilities. Another loss term is calculated as KL divergence between the sample-reference attention map and a standard mapping which connects points with same retention time in sample and reference (similar to an identity matrix). The model is trained by the cross entropy loss plus an extra divergence term on the same training set under same settings.

ADDITIONAL DEFINITIONS

[000135] As used herein a "processor" should be taken to include one or a plurality of processors used to perform a task, such as executing SAS program code, processing data for a GUI, or reading or writing from memory and includes a general purpose computer.

[000136] As used herein a "user" should be taken to include an actual person using an application locally, or accessing the application remotely using a web browser using a computer located remotely.

[000137] As used herein, a "computer" should be taken to include computers that have a hard disk drive for reading from and writing to a hard disk, a magnetic disk drive for reading from or writing to a removable magnetic disk, and an optical disk drive for reading from or writing to a removable optical disk such as a CD ROM, DVD, or other optical media. The hard disk drive , magnetic disk drive , and optical disk drive are connected to the system bus by a hard disk drive interface, a magnetic disk drive interface, and an optical drive interface, respectively. The drives and their associated computer-readable media provide nonvolatile (non-transitory) storage of computer readable instructions, data structures, programs, and other data for the computer system.

[000138] As used herein, a "database" should be taken to include a centralized or distributed database and includes a SQL database.

[000139] As used herein, the term "data" should be taken to include any information contained in machine readable memory of a file system or used in a database or other structured data, or unstructured data representation in transitory or non-transitory memory.

[000140] As herein used, an "input device" should be taken to include a keyboard, mouse, pointer, touch pad, voice recognizer or other forms of input devices connected to a computer (with or without a display) for making modifications to, exploring, accessing, or generating electronic content that may be stored in memory.

[000141] As herein used, "display" is an LCD/LED type display or other types of display devices connected to the computer and capable of rendering and receiving input using a GUI or displaying other information generated by an application running on the computer executing machine executable code.

[000142] As used herein, "module" or "component" should be taken to include a software or source code file, or set of machine-executable instructions contained, or not contained in a source code file and existing on a non-transitory or transitory memory medium. A module may exist as a stand-alone program or represent a program called by another program. A "component" is a portion of a "module" in the sense that the module calls a component to perform a task or task(s) portion of the module.

[000143] A "data repository" or "repository" as used and construed can include a database, data store, data warehouse, data lake or a combination of these things.

[000144] As used herein "memory" should be taken to include a machine-readable medium, either transient computer-readable storage medium or non-transient computer-readable storage medium. In various embodiments, the volatile (transient) portion of memory may be implemented using any suitable memory technology, such as static random access memory (SRAM), synchronous dynamic RAM or any other type of memory. While the examples of memory as shown and described may refer to a single machine-readable medium, memory should be taken to include a single medium or multiple media (e.g., a centralized or distributed database, and/or associated caches and servers, and or a variety of storage media, such as the processor registers, memories, and the storage device) that store one or more sets of machine-executable code or data. The term "machine-readable medium" shall also be taken to include any medium that is capable of storing, encoding or carrying a set of instructions for execution by the machine and that cause the machine to perform any one or more of the methodologies of the present invention, or that is capable of storing,

encoding or carrying data structures utilized by or associated with such a set of instructions. The term memory shall accordingly be taken to include tangible media, such as solid-state memories and optical and magnetic media.

[000145] As used herein “module” should be taken to include a unit of distinct functionality that can provide information to, and receive information from, other modules. Accordingly, described modules may be regarded as being communicatively coupled. Modules may also initiate communication with input or output devices, and can operate on a resource (e.g., a collection of information, such as a database). Modules may include hardware circuitry, optical components, single or multi-processor circuits, memory circuits, software program modules and objects, firmware, and combinations thereof, as appropriate for particular implementations of various embodiments. The term "module" can include an identifiable portion of code, data, or a computational object to achieve a particular function, operation, processing, or procedure.

In some examples ,set forth herein is a method for diagnosing a disease comprising

[000146] quantifying mass spectrometry transitions , wherein the transitions are m/z intensity value peaks characteristic of glycopeptides or peptides comprising an amino acid sequence, and analyzing the quantification using a model to determine a probability that the sample is or isn’t within a disease classification based on the quantification. In some of these examples, the transitions, glycopeptides, peptides, amino acid sequences, or any combination thereof may be those set forth in International PCT Patent Application No. PCT/US2020/0162861, filed January 31, 2020.

[000147] The above description of illustrated embodiments of the invention, including what is described in the Abstract, is not intended to be exhaustive or to limit the invention to the precise forms disclosed. While specific embodiments of, and examples for, the invention are described herein for illustrative purposes, various modifications are possible within the scope of the invention, as those skilled in the relevant art will recognize.

[000148] These modifications can be made to the invention in light of the above detailed description. The terms used in claims should not be construed to limit the invention to the specific embodiments disclosed in the specification.

WHAT IS CLAIMED IS:

1. A method of training a model to identify mass spectrometry peaks, comprising:

using a plurality of labeled sequential data, each of the plurality of labeled sequential data comprising a peak start and a peak stop label representing retention time values characteristic of glycopeptides or peptides, or fragments thereof, and

training at least one machine learning model using the labeled sequential data, wherein the trained model is adapted for identifying mass spectrometry start and stop peaks indicating the presence of one or more glycopeptides or peptides, or a fragment thereof, present in a biologic sample.

2. The method of claim 1, further comprising quantifying the presence of the one or more glycopeptides or peptides by integration over time between the mass spectrometry start and stop peaks from the biologic sample.

3. The method of claim 1, wherein the machine learning model is a recurrent neural network model.

4. The method of claim 3, wherein the recurrent neural network is a bi-directional LSTM recurrent neural network model with attention layer.

5. A method comprising computing an abundance for a glycopeptide or peptide using a machine learning model trained by the process of claim 1.

6. A system, comprising

a data repository of labeled sequential data;

one or more processors coupled to the repository;

machine executable code, residing in non-transitory memory accessible to the one or more processors, that when executed by the one or more processors performs the method of any of claims 1 – 5.

7. A method of identifying mass spectrometry peaks, comprising:

providing at least one trained neural network model that produces an output as a function of input parameters, wherein values for the input parameters comprise a sequence of

retention time values for a glycopeptide having a m/z (mass to charge) ratio; and

providing as values for the input parameters mass spectrometry (MS) data comprising mass-to-charge (m/z) ratios over the sequence of retention time values from a mass spectrometry of a biologic sample;

wherein the output comprises a location for one or more peak start and peak stop retention time values indicating the presence of one or more analytes, or a fragment thereof, in the biologic sample.

8. The method of claim 7, further comprising quantifying the presence of the one or more glycopeptides or peptides by summing or integration over retention time values between the mass spectrometry start and stop peaks from the biologic sample.

9. The method of Claim 8, wherein the quantifying step includes applying a weighted smoothing function over the predicted start and stop peaks.

10. A system, comprising

a data repository of mass spectrometry (MS) data;

one or more processors coupled to the repository;

machine executable code, residing in non-transitory memory accessible to the one or more processors, that when executed by the one or more processors performs the method of any of claims 7 – 9.

11. A method, comprising:

featurizing data from raw MS data, wherein the raw MS data comprises a plurality of analog samples, each representing mass/charge intensity, the featurizing comprising, for each of the analog samples, the steps of

centering the sample within a retention time window,

discretizing the sample into a sequence of points representing intensity values,

standardizing the intensity values, and

assigning labels to points among the sequence of points corresponding to peak start and peak stop times to produce a labeled sequence of points;

wherein the labeled sequence of points are configured for training a machine learning model to predict an abundance in unseen MS data.

12. The method of Claim 11, wherein the centering comprises positioning an apex of a curve at a center of the retention window.

13. The method of Claim 11, wherein the standardizing the intensity values excludes intensity values greater than a threshold amount.

14. The method according to any of Claims 13, wherein the threshold amount is about 500.

15. A system, comprising

a data repository of raw mass spectrometry (MS) data comprising a plurality of analog samples;

one or more processors coupled to the repository;

machine executable code, residing in non-transitory memory accessible to the one or more processors, that when executed by the one or more processors performs the method of any of Claims 11 – 14.

16. A method, comprising

computing a distribution of probabilities over a sequence of points representing mass spectrometry (MS) data for a given m/z ratio and retention time period, the probabilities representing a likely peak start or peak stop retention time, and

computing a statistic for the abundance using the distribution of probabilities, wherein the statistic is indicative of the presence of a biologic sample in the MS data.

17. The method of Claim 16, wherein the statistic is a mean abundance computed as a weighted sum of integrals using all possible pairs of start and end points for evaluating the integral.

18. The method of any of Claims 1, 7 or 11, wherein the machine learning model is a sequential model (MODEL A) or a reference-based model (MODEL B).

19. A method, comprising:

using a plurality of labeled sequential data, and

training at least one machine learning model using the labeled sequential data,

wherein the trained model is adapted for identifying a time period or frequency range where a start and stop peak of an intensity occur in an unseen sample.

20. The method of Claim 19, wherein the start and stop peak is contained in a noise to signal transition time period.

21. The method of Claim 19, wherein the machine learning model is a recurrent neural network with an attention layer.

22. The method of Claim 19, wherein the sample represents a physical thing such that a summation or integration over time or frequency from the start peak to the stop peak of the intensity quantifies the physical thing.

23. The method of Claim 19, wherein the sample is a sequential series of mass to charge intensities over a specified time period and the summation or integration represents the abundance of the physical thing.

24. A method, comprising:

providing at least one trained neural network model that produces an output as a function of input parameters, wherein values for the input parameters are a temporal or spectral sequence of intensity values representing a quantity; and

providing values for each of the input parameters, the values representing an unseen sequence of intensities for a quantity;

wherein the output comprises one or more peak start and peak stop times or frequencies, which are used to quantify the quantity.

25. The method of Claim 24, wherein the quantity is quantified by summing or integration over time or frequency between the peak start and peak stop times or frequencies, respectively.

26. The method of Claim 24 or Claim 22, wherein the intensities are obtained from a process selected from the set of mass spectrometry, MRM, liquid or gas chromatography, x-ray diffraction, Raman spectroscopy, UV-VIS spectroscopy, fluorimetry spectroscopy or phosphorimetry spectroscopy, Nuclear Magnetic Resonance (NMR) spectroscopy, or

Magnetic Resonance Imaging (MRI).

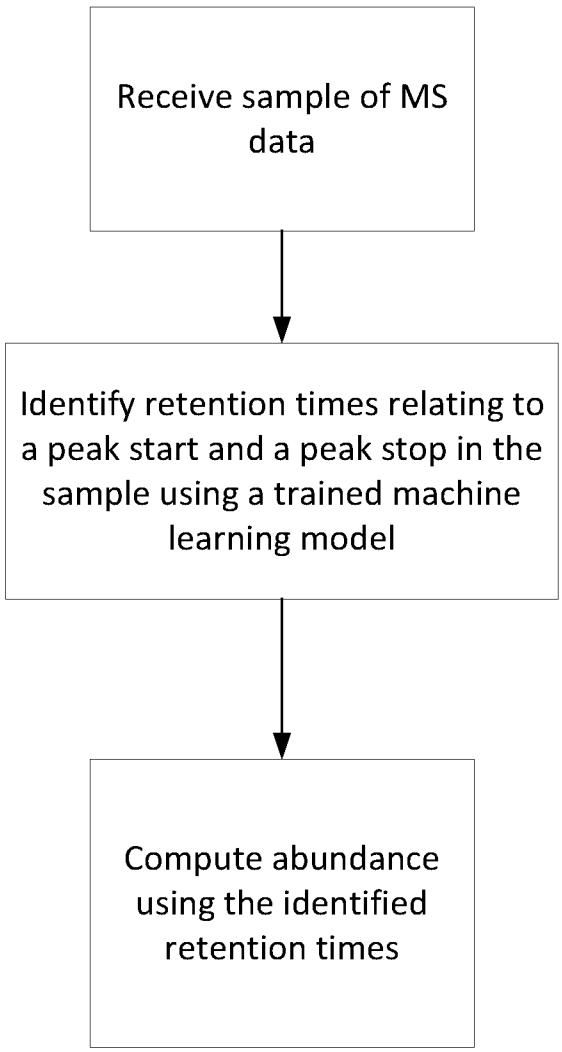

Receive sample of MS data
Identify retention times relating to a peak start and a peak stop in the sample using a trained machine learning model
Compute abundance using the identified retention times

FIG. 1

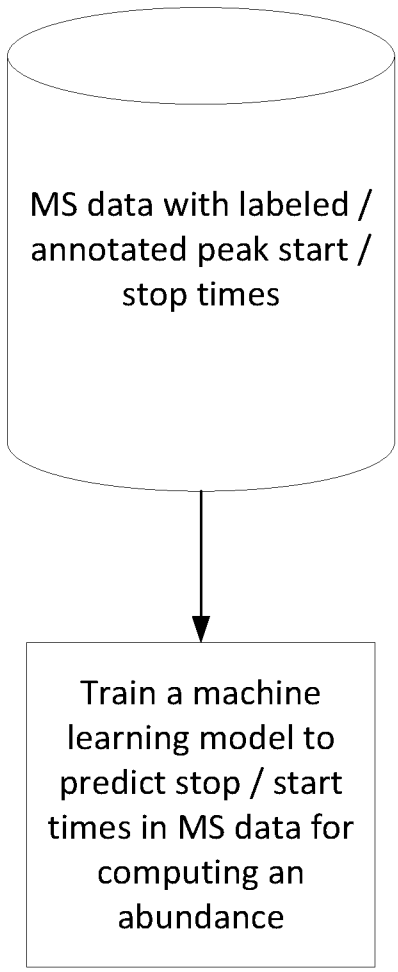

MS data with labeled / annotated peak start / stop times
Train a machine learning model to predict stop / start times in MS data for computing an abundance

FIG. 2

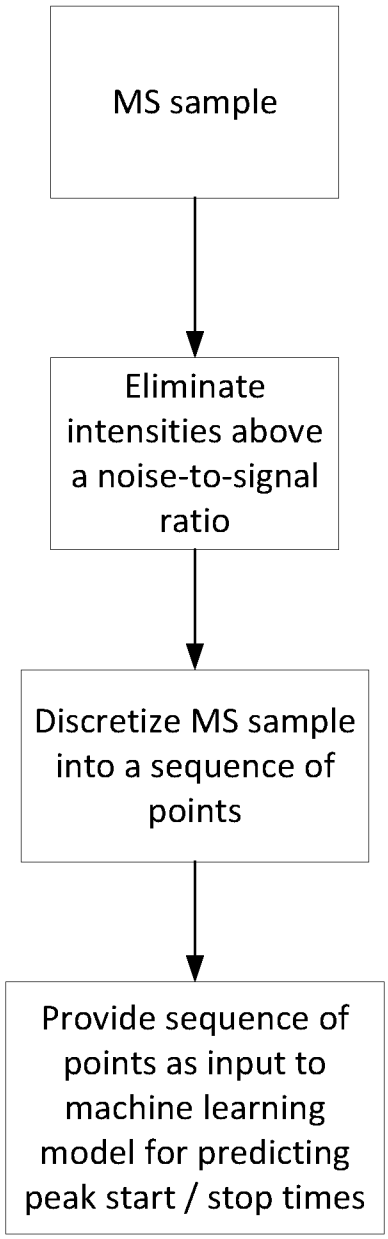

MS sample
Eliminate intensities above a noise-to-signal ratio
Discretize MS sample into a sequence of points
Provide sequence of points as input to machine learning model for predicting peak start / stop times

FIG. 3

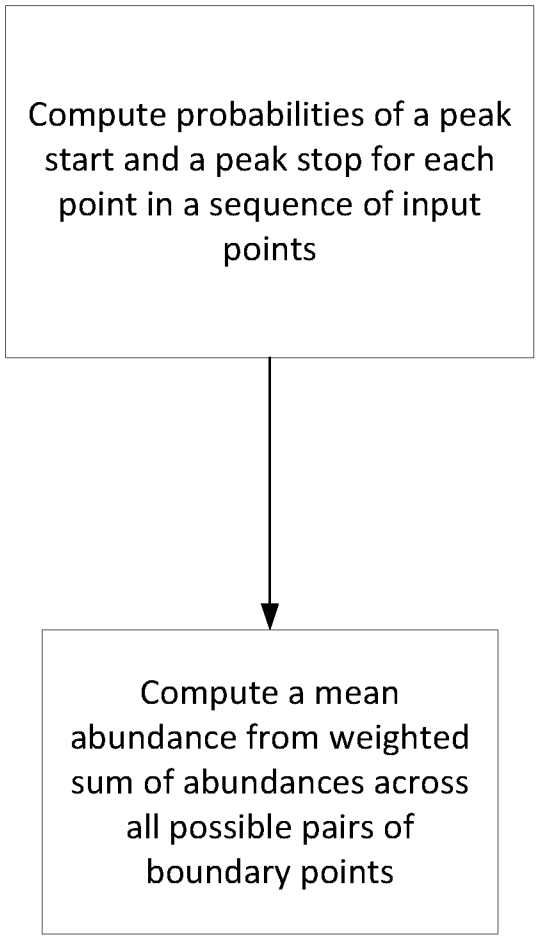

Compute probabilities of a peak start and a peak stop for each point in a sequence of input points
Compute a mean abundance from weighted sum of abundances across all possible pairs of boundary points

FIG. 4

FIG. 5A

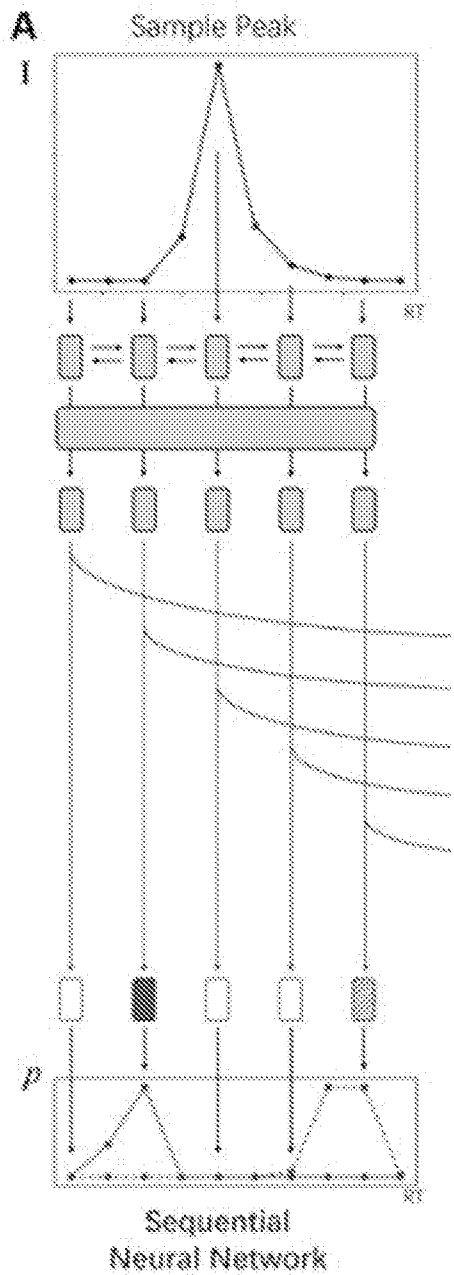

A
Sample Peak
I
Sequential
Neural Network

FIG. 5B

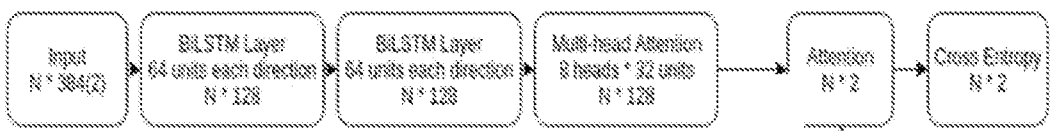

Input
BiLSTM Layer
64 units each direction
BiLSTM Layer
64 units each direction
Multi-head Attention
Attention
Cross Entropy

FIG. 6A

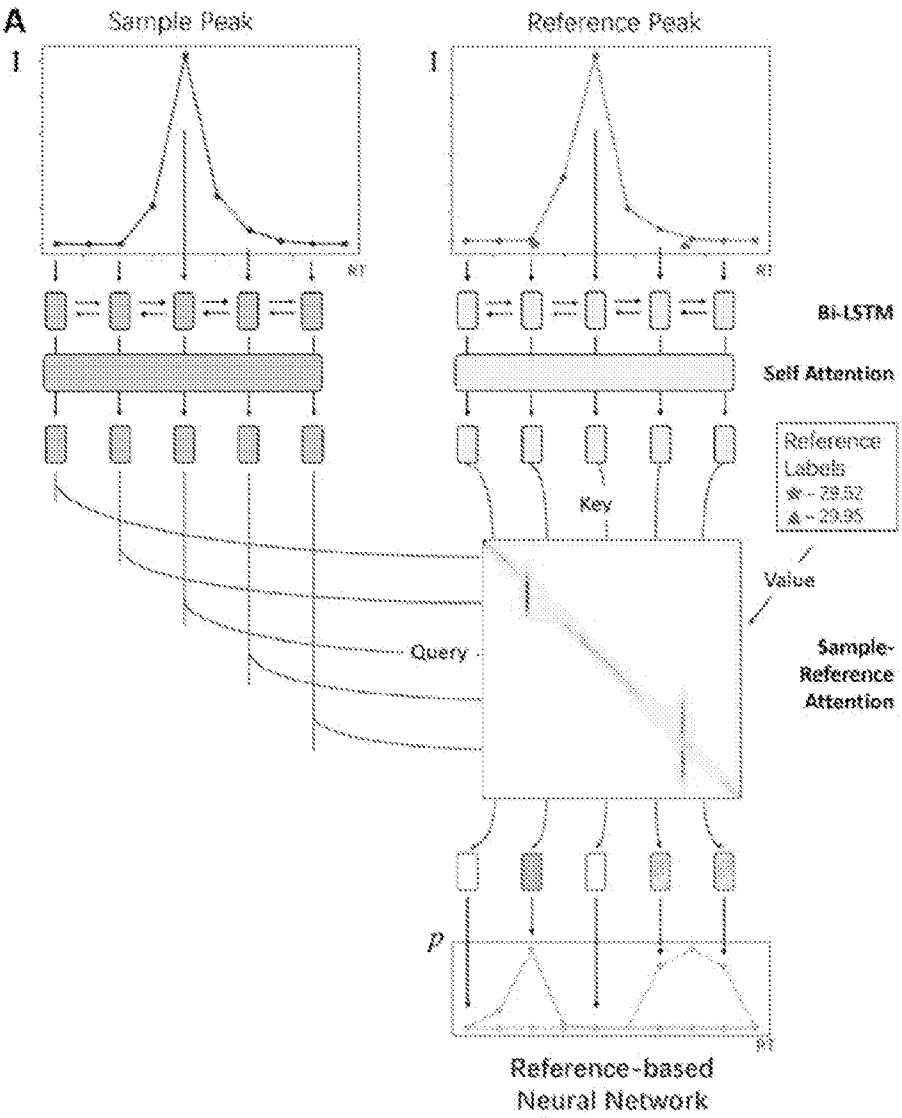
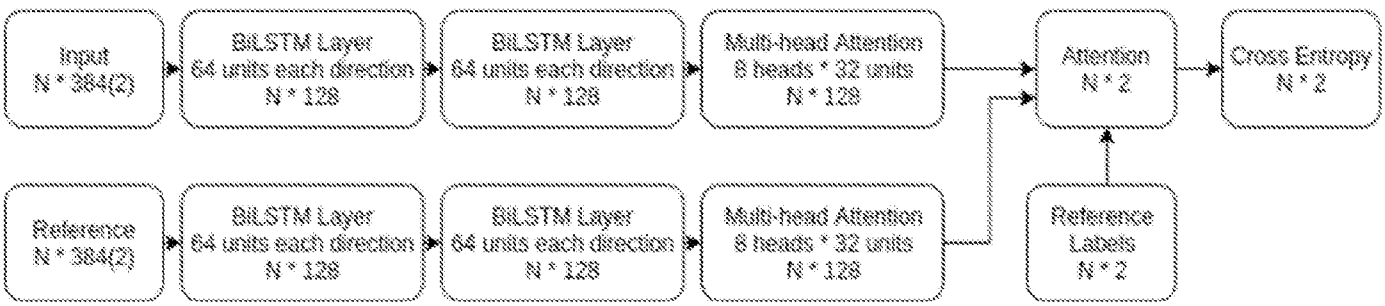

A
Sample Peak
Reference Peak
I
I
Bi-LSTM
Self Attention
Reference
Labels
Key
Value
Query
Sample-
Reference
Attention
P
Reference-based
Neural Network

FIG. 6B


Input
N * 384(2)
BiLSTM Layer
64 units each direction
N * 128
BiLSTM Layer
64 units each direction
N * 128
Multi-head Attention
8 heads * 32 units
N * 128
Attention
N * 2
Cross Entropy
N * 2
Reference
N * 384(2)
BiLSTM Layer
64 units each direction
N * 128
BiLSTM Layer
64 units each direction
N * 128
Multi-head Attention
8 heads * 32 units
N * 128
Reference
Labels
N * 2

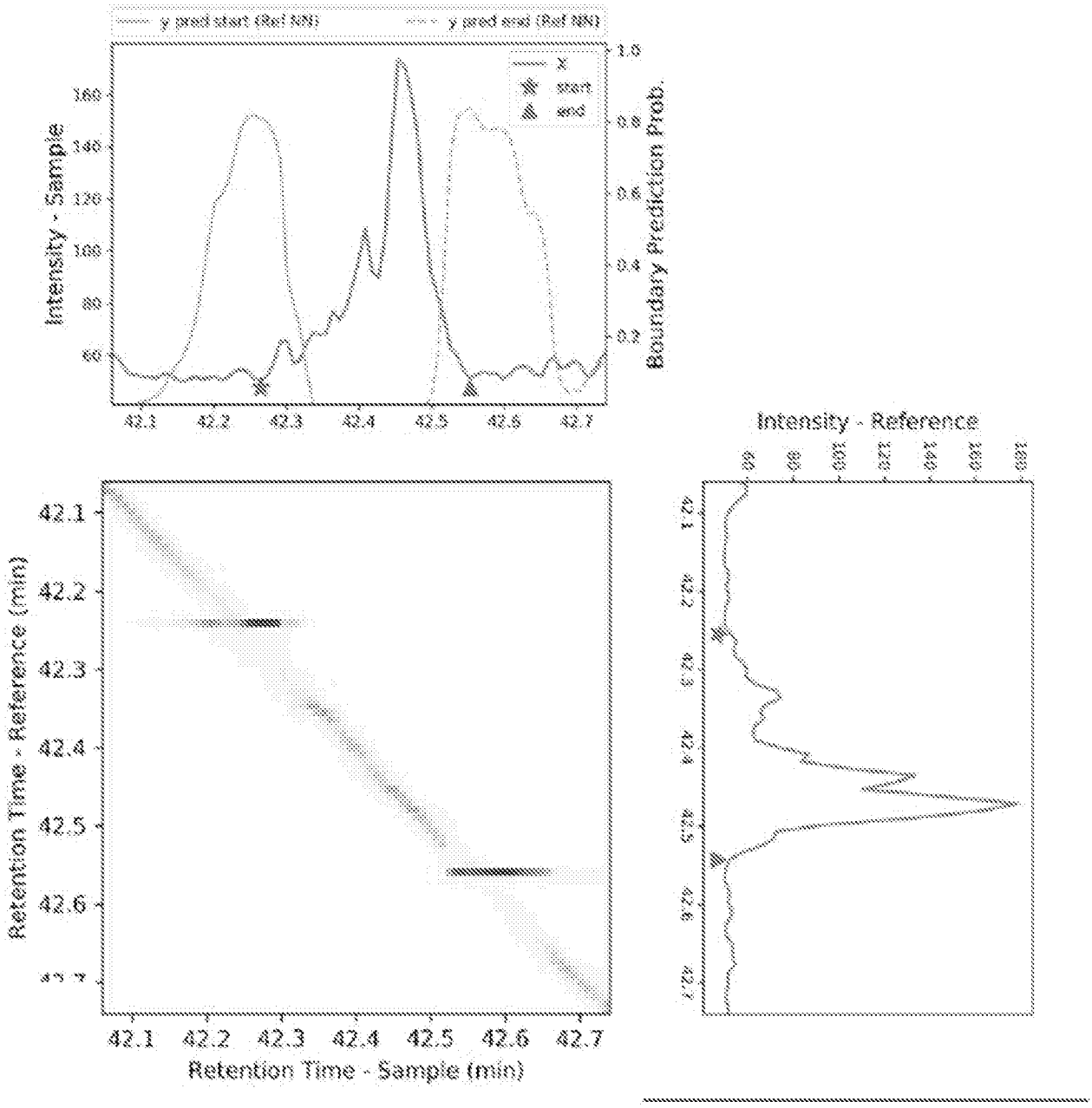

y pred start (Ref NN)
y pred end (Ref NN)
x
start
end
Intensity - Sample
Boundary Prediction Prob.
Intensity - Reference
Retention Time - Reference (min)
Retention Time - Sample (min)
42.1
42.2
42.3
42.4
42.5
42.6
42.7

FIG. 6C

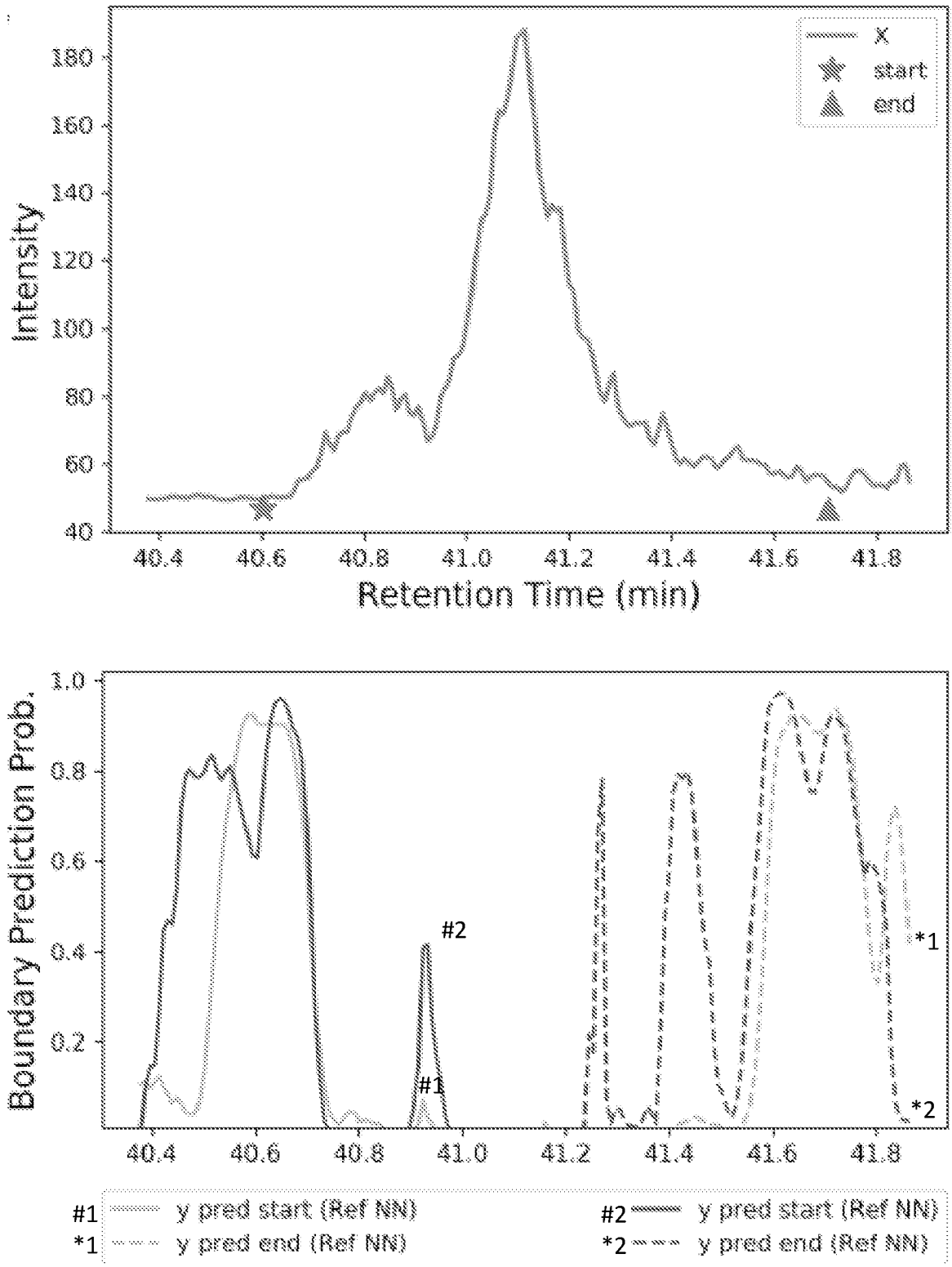

X
start
end
Intensity
180
160
140
120
100
80
60
40
40.4
40.6
40.8
41.0
41.2
41.4
41.6
41.8
Retention Time (min)
Boundary Prediction Prob.
1.0
0.8
0.6
0.4
0.2
#1
#2
*1
*2
#1 y pred start (Ref NN)
*1 y pred end (Ref NN)
#2 y pred start (Ref NN)
*2 y pred end (Ref NN)

FIG. 7A

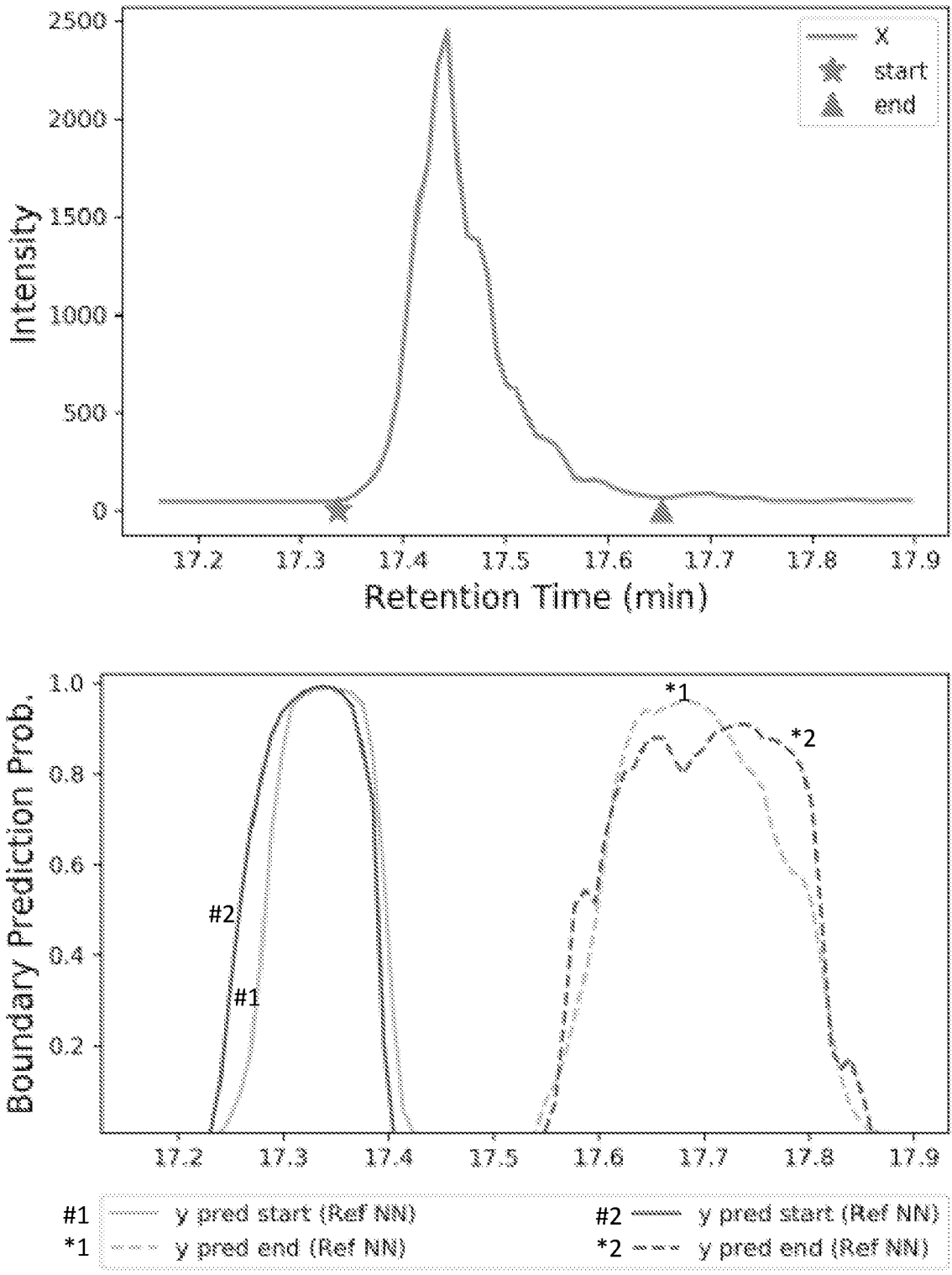

2500
2000
1500
1000
500
0
Intensity
X
start
end
17.2
17.3
17.4
17.5
17.6
17.7
17.8
17.9
Retention Time (min)
1.0
0.8
0.6
0.4
0.2
Boundary Prediction Prob.
#1
#2
*1
*2
#1 y pred start (Ref NN)
*1 y pred end (Ref NN)
#2 y pred start (Ref NN)
*2 y pred end (Ref NN)

FIG. 7B

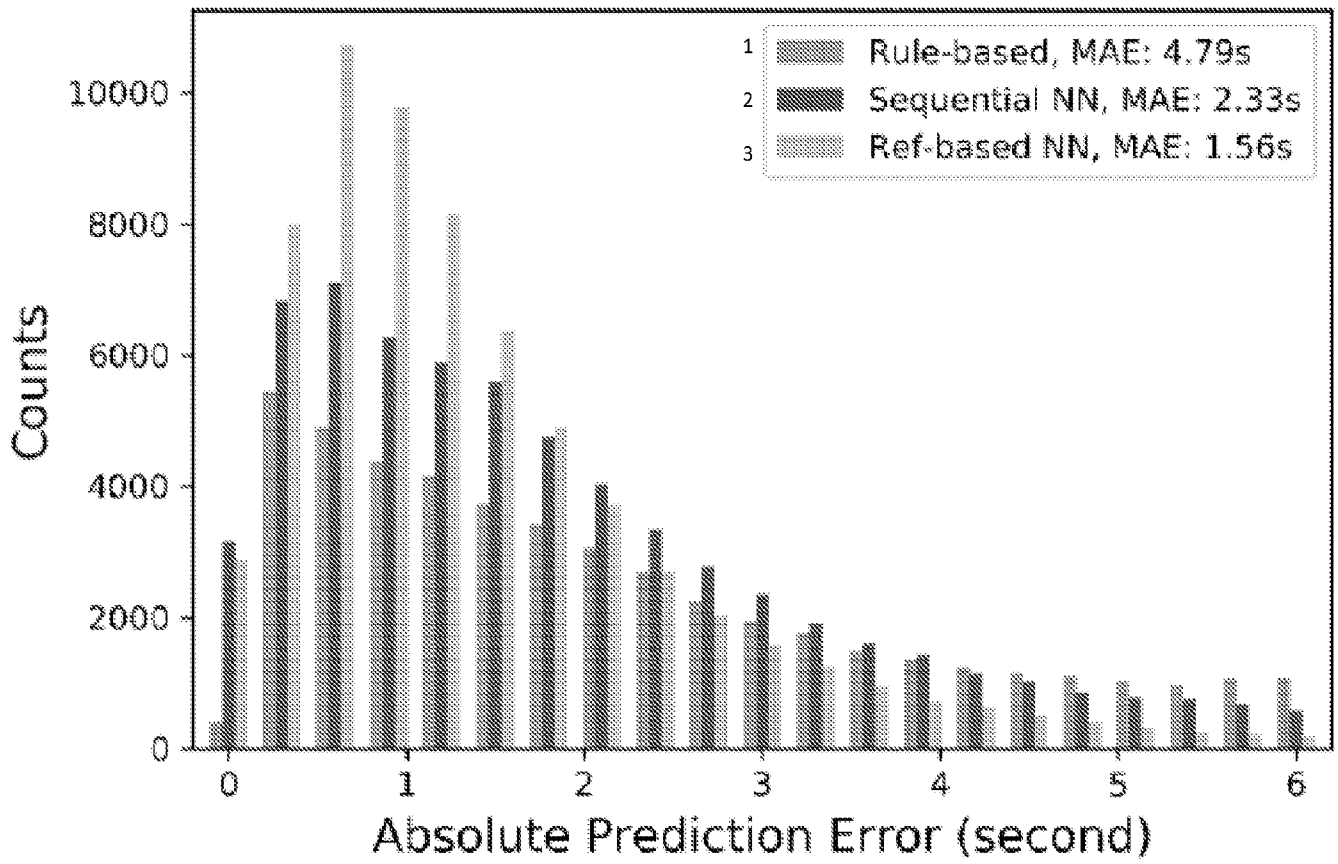

1 Rule-based, MAE: 4.79s
2 Sequential NN, MAE: 2.33s
3 Ref-based NN, MAE: 1.56s
Counts
10000
8000
6000
4000
2000
0
0
1
2
3
4
5
6
Absolute Prediction Error (second)

FIG. 8A

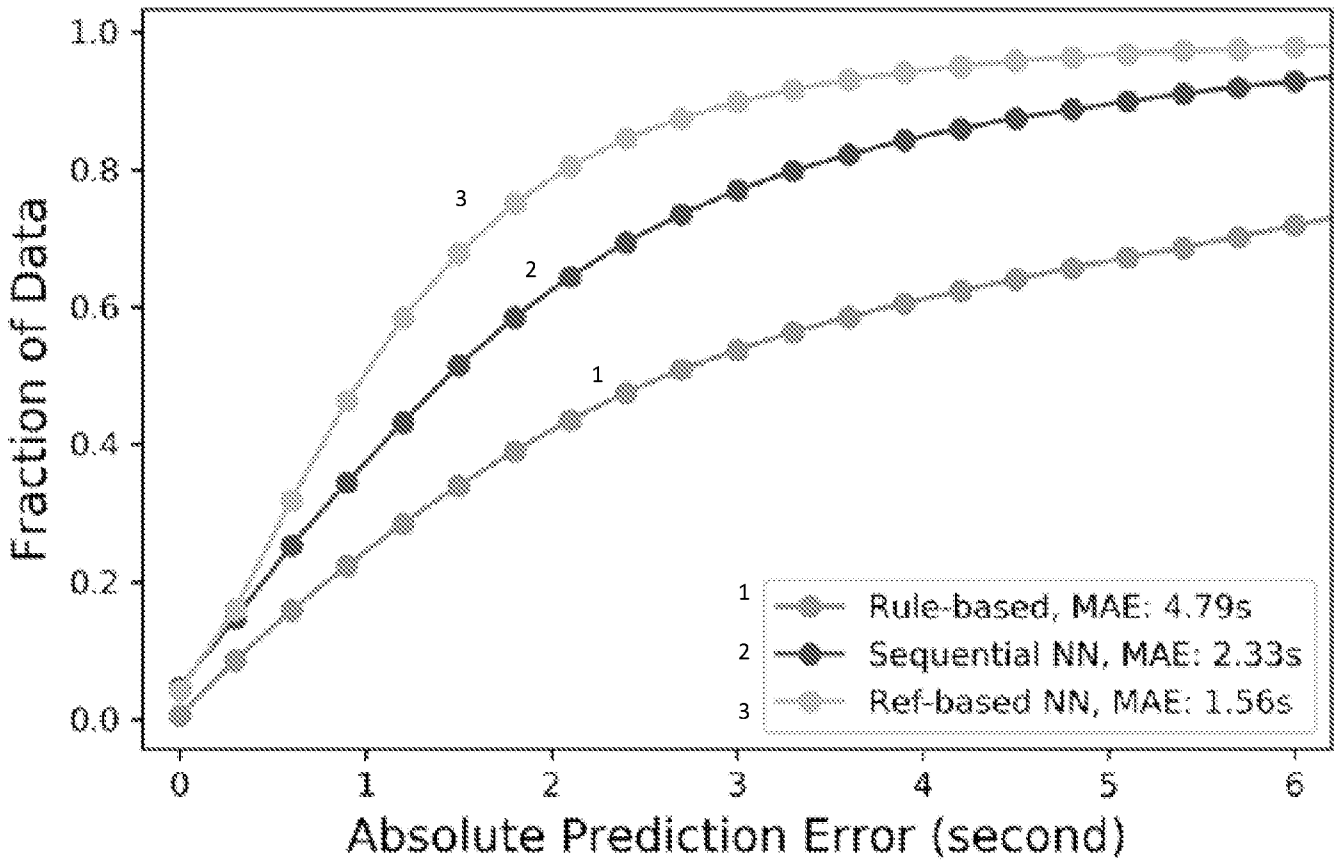

Fraction of Data
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
3
4
5
6
Absolute Prediction Error (second)
1
2
3
Rule-based, MAE: 4.79s
Sequential NN, MAE: 2.33s
Ref-based NN, MAE: 1.56s

FIG. 8B

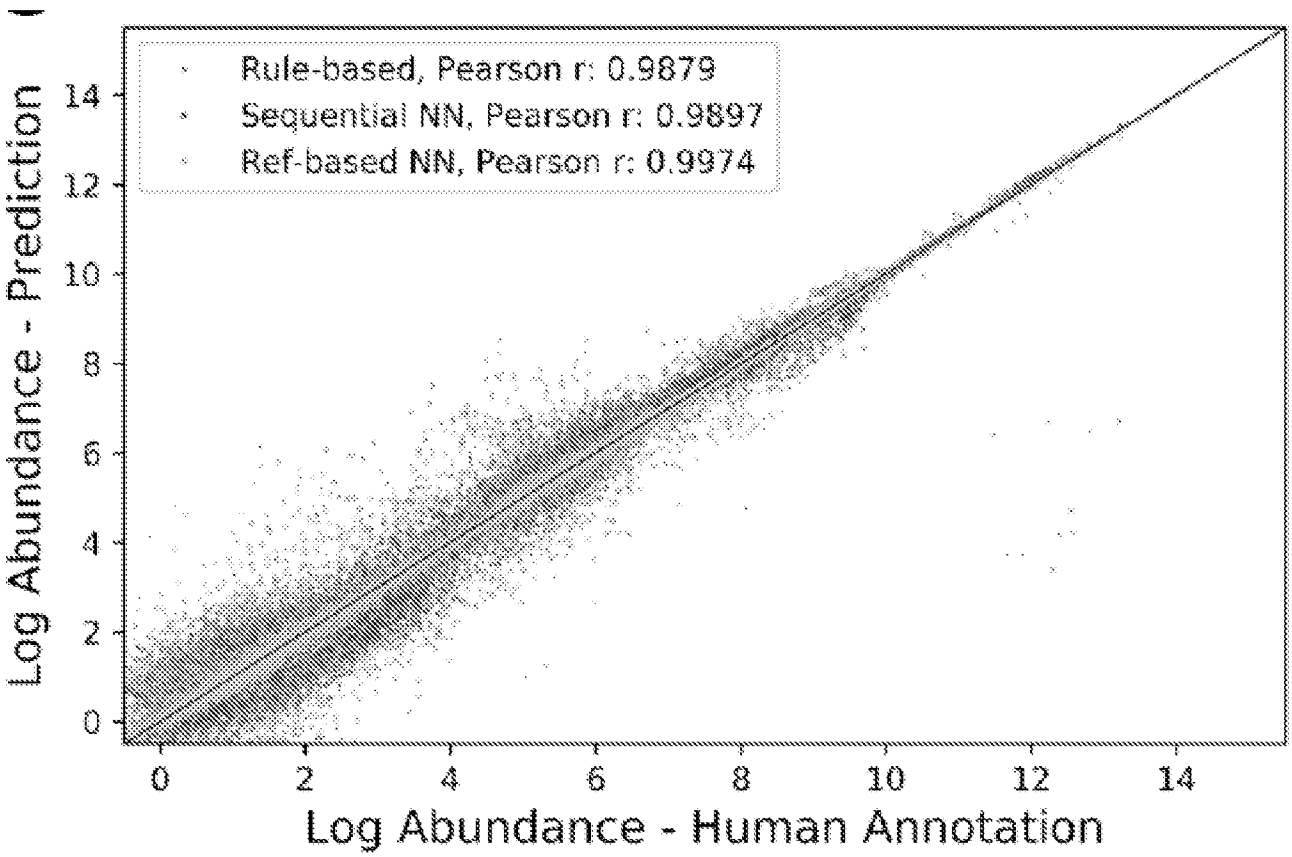

Rule-based, Pearson r: 0.9879
Sequential NN, Pearson r: 0.9897
Ref-based NN, Pearson r: 0.9974
Log Abundance - Prediction
Log Abundance - Human Annotation
0
2
4
6
8
10
12
14

FIG. 8C

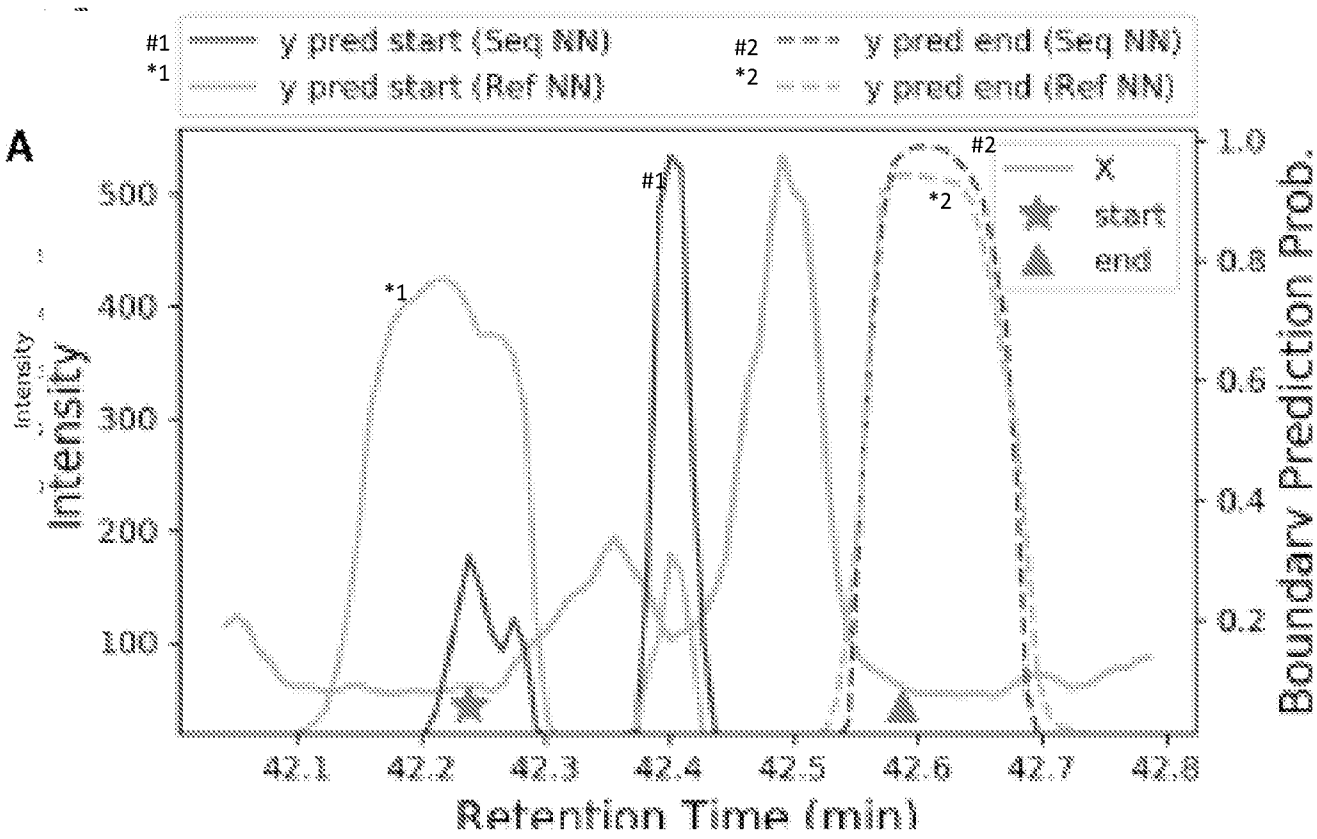

A
#1
*1
y pred start (Seq NN)
y pred start (Ref NN)
#2
*2
y pred end (Seq NN)
y pred end (Ref NN)
X
start
end
Intensity
Boundary Prediction Prob.
Retention Time (min)
100
200
300
400
500
0.2
0.4
0.6
0.8
1.0
42.1
42.2
42.3
42.4
42.5
42.6
42.7
42.8

FIG. 9A

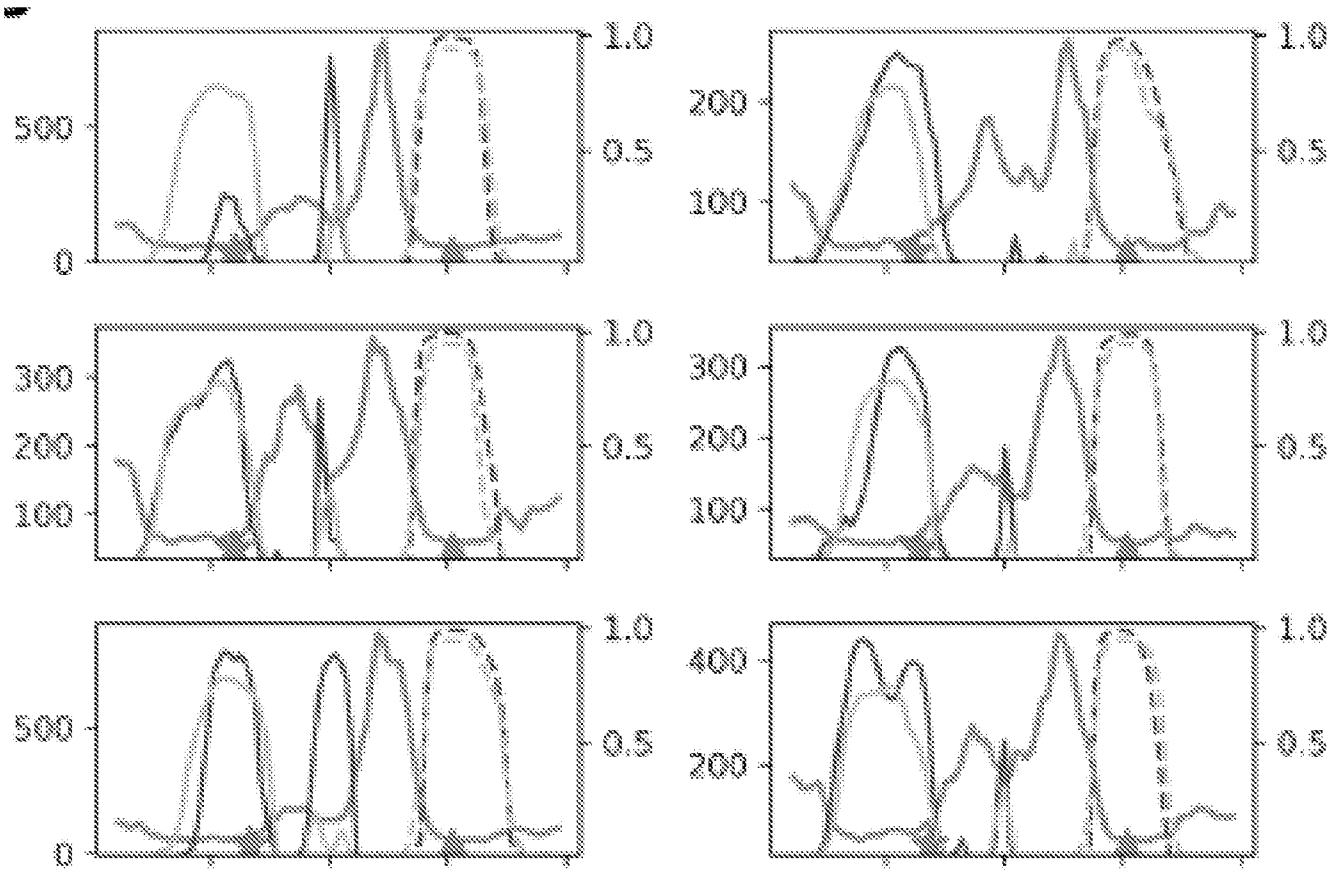

500
0
1.0
0.5
200
100
1.0
0.5
300
200
100
1.0
0.5
300
200
100
1.0
0.5
500
0
1.0
0.5
400
200
1.0
0.5

FIG. 9B

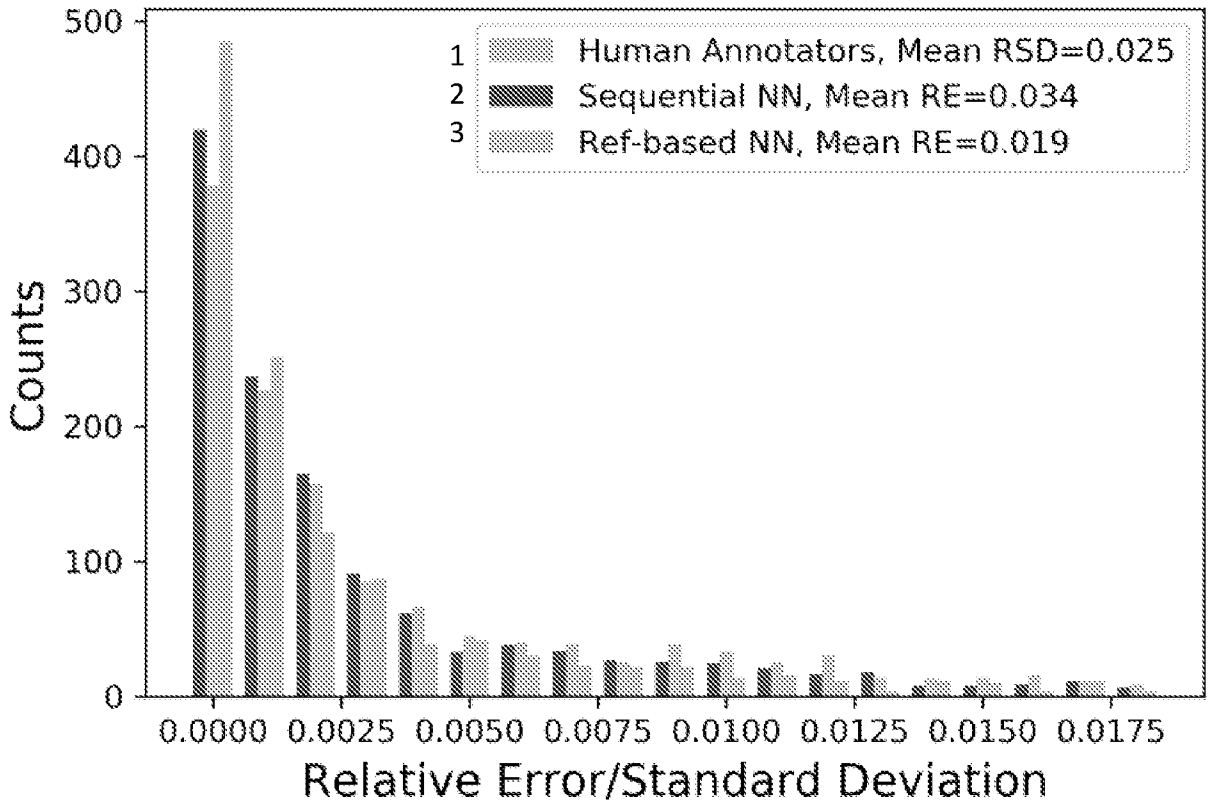

500
400
300
200
100
0
Counts
1
2
3
Human Annotators, Mean RSD=0.025
Sequential NN, Mean RE=0.034
Ref-based NN, Mean RE=0.019
0.0000
0.0025
0.0050
0.0075
0.0100
0.0125
0.0150
0.0175
Relative Error/Standard Deviation

FIG. 10

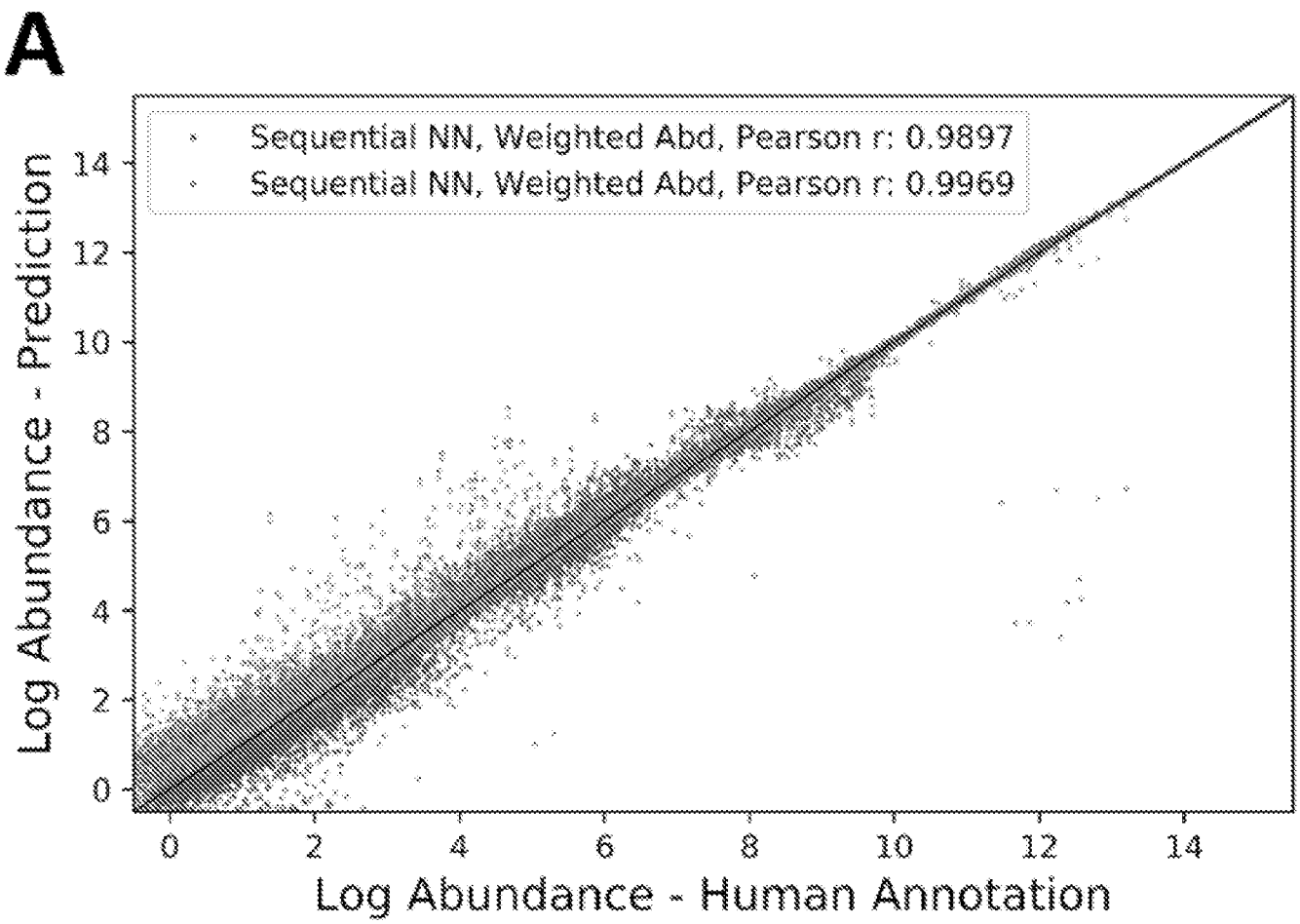

A
Sequential NN, Weighted Abd, Pearson r: 0.9897
Sequential NN, Weighted Abd, Pearson r: 0.9969
Log Abundance - Prediction
Log Abundance - Human Annotation
0
2
4
6
8
10
12
14

FIG. 11A

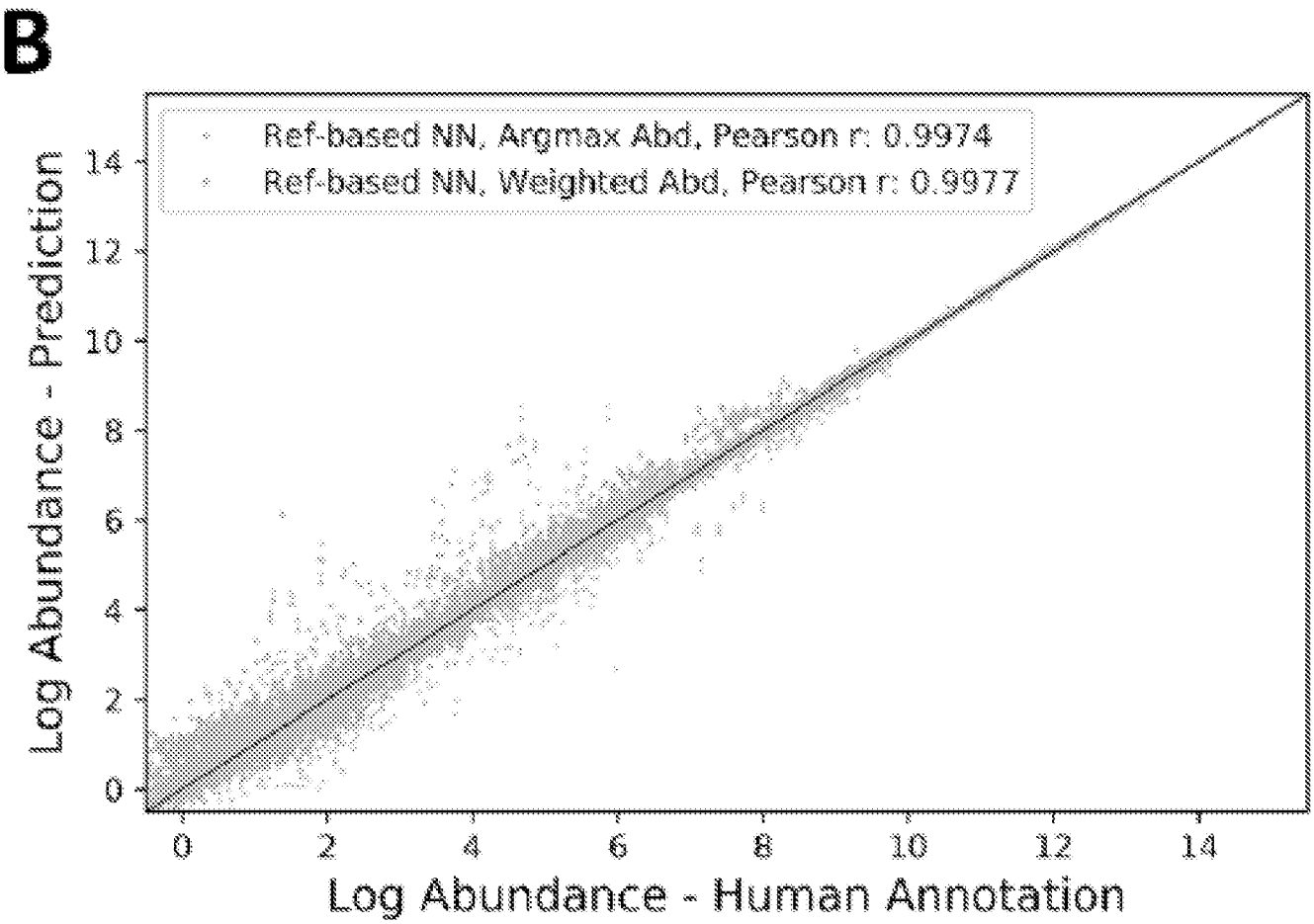

B
Ref-based NN, Argmax Abd, Pearson r: 0.9974
Ref-based NN, Weighted Abd, Pearson r: 0.9977
Log Abundance - Prediction
Log Abundance - Human Annotation
0
2
4
6
8
10
12
14

FIG. 11B

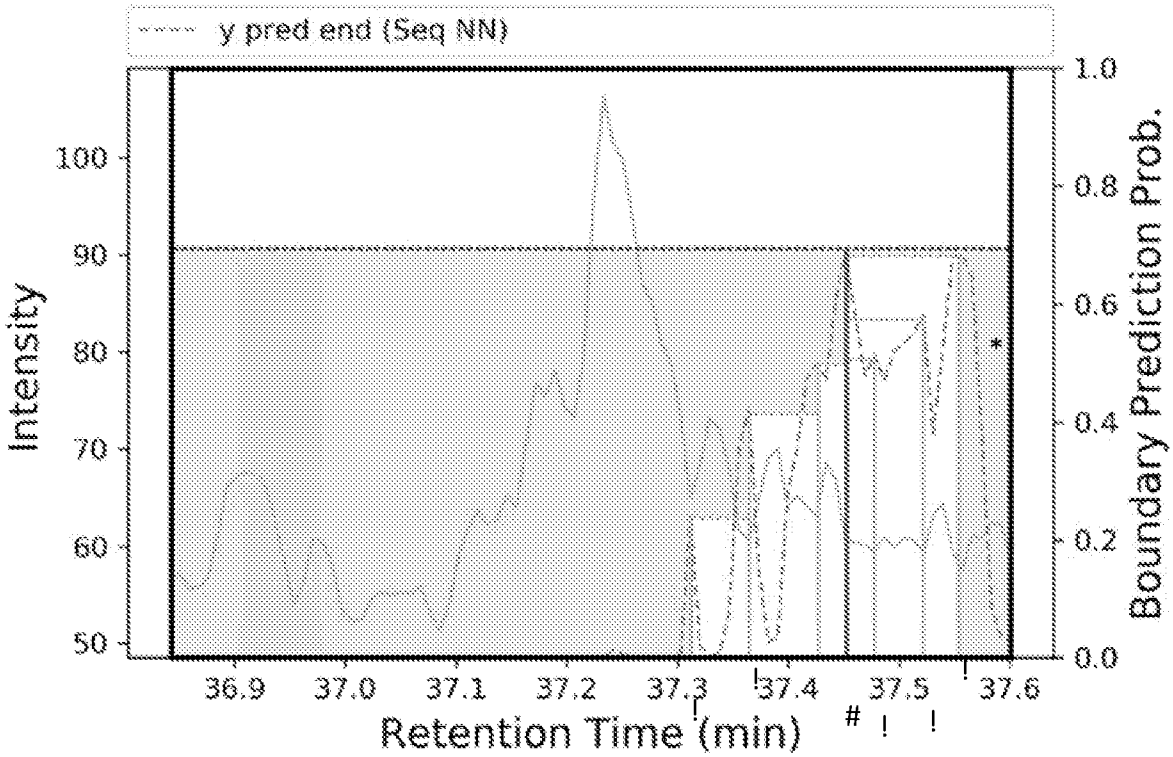

y pred end (Seq NN)
Intensity
100
90
80
70
60
50
Boundary Prediction Prob.
1.0
0.8
0.6
0.4
0.2
0.0
36.9
37.0
37.1
37.2
37.3
37.4
37.5
37.6
Retention Time (min)

FIG. 12

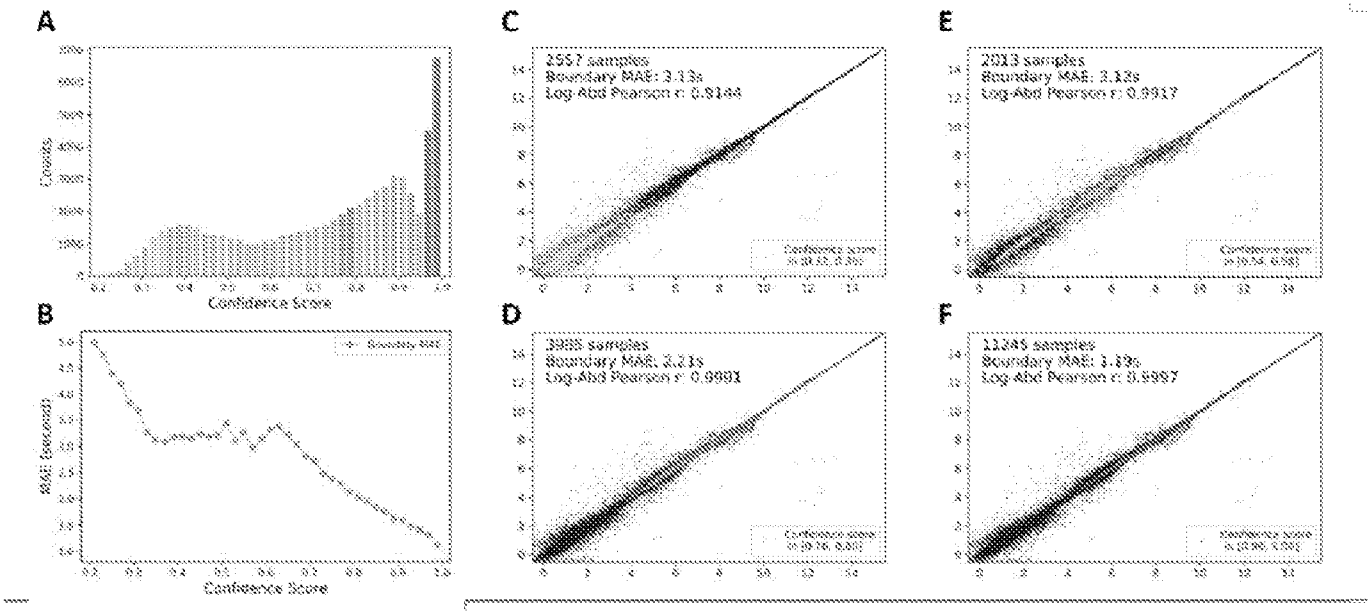


FIG. 13A, 13B, 13C, 13D and 13E

A



FIG. 14A

B

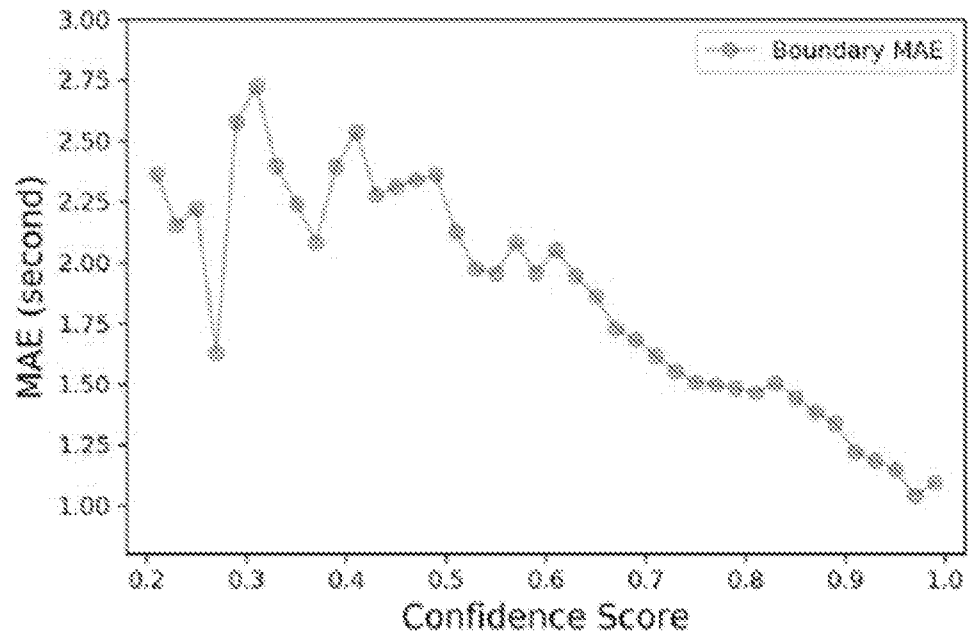


FIG. 14B

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2020/025502

A. CLASSIFICATION OF SUBJECT MATTER
INV. G16B40/20
ADD. G16H10/40 G16H50/20 G16H50/70

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G16B G16H

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	SHADI TOGHI ESHGHI ET AL: "Quality assessment and interference detection in targeted mass spectrometry data using machine learning", CLINICAL PROTEOMICS, SPRINGER, US, vol. 15, no. 1, 6 October 2018 (2018-10-06), pages 1-13, XP021261354, ISSN: 1542-6416, DOI: 10.1186/S12014-018-9209-X abstract p. 1/13 to 2/13 section "Background" p. 4/13 section "Engineering QC features for peak quality assessment" and section "Creating the training data sets" p. 4/13 to 5/13, section "Building the predictive quality assessment model" ----- -/--	1-4,6

[X] Further documents are listed in the continuation of Box C. [ ] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
19 May 2020	20/07/2020

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Heidrich, Alexander

1

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GALLANT S R ET AL: "Deconvolution of overlapping chromatographic peaks using a cerebellar model arithmetic computer neural network", CHEMOMETRICS AND INTELLIGENT LABORATORY SYSTEMS, ELSEVIER SCIENCE PUBLISHERS B.V. AMSTERDAM, NL, vol. 18, no. 1, 1 January 1993 (1993-01-01), pages 41-57, XP026654534, ISSN: 0169-7439, DOI: 10.1016/0169-7439(93)80043-H [retrieved on 1993-01-01] abstract p. 44, section "Chromatographic signal generation"; figures 1, 2	7-10, 19-26
X	Vadim Demichev ET AL: "DIA-NN: Neural networks and interference correction enable deep coverage in high-throughput proteomics", bioRxiv, 25 October 2018 (2018-10-25), XP055696494, DOI: 10.1101/282699 Retrieved from the Internet: URL:https://www.biorxiv.org/content/10.1101/282699v2.full.pdf [retrieved on 2020-05-18] the whole document	7,10
X	TRAN NGOC HIEU ET AL: "Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry", NATURE METHODS, NATURE PUB. GROUP, NEW YORK, vol. 16, no. 1, 20 December 2018 (2018-12-20), pages 63-66, XP036661431, ISSN: 1548-7091, DOI: 10.1038/S41592-018-0260-3 [retrieved on 2018-12-20] the whole document	7,10
X	FATEMA TUZ ZOHORA ET AL: "DeepIso: A Deep Learning Model for Peptide Feature Detection", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 9 December 2017 (2017-12-09), XP081203333, the whole document	7,10

1

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2020/025502

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

see additional sheet

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☒ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

1-4, 6-10, 19-26

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

This International Searching Authority found multiple (groups of) inventions in this international application, as follows:

1. claims: 1-4, 6-10, 19-26

   Deconvolution of overlapping intensity peaks in spectrometry data

   ---

2. claims: 5, 11-18

   Quantification of mass spectrometry peaks
